Universitá di Pisa

Doctoral Thesis

Study and development of methods for the characterization of voice in emotional expressions and in mental disorders

Author:
Andrea Guidi

Supervisor:
Prof. Enzo Pasquale Scilingo,
Eng. Nicola Vanello

Doctoral School of Engineering "L. Da Vinci"

Ph.D. Programme in Automation, Robotics, Bioengineering

Ph.D. Cycle Number: XXVII

December 2015

Words mean more than what is set down on paper. It takes the human voice to infuse them with shades of deeper meaning.

Maya Angelou

But some emotions don't make a lot of noise.

Ernest Hemingway

UNIVERSITÁ DI PISA

Abstract

Faculty of Engineering
Doctoral School of Engineering "L. Da Vinci"

Doctor of Philosophy

Study and development of methods for the characterization of voice in emotional expressions and in mental disorders

by Andrea GUIDI

In this thesis, voice signal is investigated with the aim of characterizing subjects' emotional state and patients' mood state. Speech signal is a nonstationary signal that is generated by a complex phenomenon that is influenced by the autonomic and somatic nervous systems, through the modulation of breathing activity, vocal muscles tension, salivation and mucus secretion. Generally, phonatory system can be modelled according to the so-called source-filter theory of speech production. In this model the source is represented by pulsatile airflow or turbulent airflow generated by the modulating action of the vocal folds. In fact, by means of their closing and opening motion, vocal folds are able to modulate the airflow coming from the lungs. Such a source, generated by the vocal folds, is hence filtered according to the resonance characteristics of the supraglottal vocal tract. Its resonances depend on its size and shape and generally they are continuously modified to allow the emission of specific sound targets.

Eckman stated that moods are emotional feelings lasting for an extended period of time, while emotions are temporary feelings that tend to come and go quite quickly. Speech can be usefully investigated to give a characterization of the emotional and/or mood state of the speakers. Many studies have been conducted to characterize both of them. Emotions can be studied by using several kinds of database where emotions can be natural, acted or induced. Moods usually are investigated in relation with some mental disorder. Especially, bipolar disorder is characterized by a great variability of moods, since bipolar patients experience sudden and sometimes extreme mood swings. Notwithstanding, in the literature the major efforts were made on the studies of patients affected by depression.

Generally, speech-related features can be thought as divided into three main categories. The first category is aiming the investigation of the prosodic dynamics of the speech. For this purpose, perceived rhythm, stress, intonation, pitch, speaking rate, and loudness are some of the possible cues that can be studied. The second one is related to the source of voice production and to the airflow streaming from the lungs through the glottis. Source features are also investigated to obtain information about voice quality, i.e. the auditory perception of the modification of vocal fold vibration and vocal tract shape. Finally, the third category is related to the spectral analysis of the speech signals.

Voice signal is a nonstationary signal that can be considered stationary if it is analysed over a sufficiently short period of time. When this kind of signal is investigated, it is important to perform a proper detection and segmentation of the voiced segments. The minimization of the rate of segment mislabelling is mandatory to achieve reliable estimates of the investigated features. Moreover, since voice signal is characterized by an high intra- and inter-day variability, it is crucial detecting carefully the features. For this purpose, features should be able to highlight statistically significant differences related to the emotion or mood state transitions, while they would show a robust behaviour with regard to the natural daily variability. In this frame, when a signal is recorded, it is fundamental to take into account also the environmental condition. Environmental noise and/or reverberation can severely alter the acquired signal, resulting in misleading results and conclusions.

In this thesis, speech signals were investigated to recognize/characterize emotional states in actors and mood states in patients affected by bipolar disease. The investigation was performed at different levels of description. Micro-prosodic and higher level phenomena were studied. Small changes of the glottal cycle related to emotion and mood were observed in the initial investigations. Then, global prosodic and vocal quality studies were conducted later. More in details, four different methods related to three different description levels have been investigated in this study. The first method is focussed on *vocal features* (lower description level) and is concerning the investigation of glottal features. These features are: mean and standard deviation of fundamental frequency and jitter. Then, two methods were focussed on *prosodic features* (mid description level). The first one took into account a prosodic analysis within every voiced segment, while the second considered globally the whole prosodic behaviour of the speech. Finally, the last method aimed at investigating the *voice quality* (higher description level) by means of the Long-Term Average Spectrum of Voice.

Synthetic datasets, healthy control subjects, and a neutral database providing both audio and electroglottographic (EGG) recordings (CMU Arctic Database) were used to test the developed algorithms. Emotional studies were conducted on the German

Emotional database, formed by actors playing different emotions. At the end, the mood investigation was performed on a database of audio samples acquired on bipolar patients. The patients were enrolled within the PSYCHE European project and performed two different vocal tasks: text reading and free image commenting.

Concerning the German Emotional database, the obtained results showed that the proposed and developed methods were able to highlight statistically significant differences among emotional speeches according to the arousal level of the acted emotion: the more the subjects are aroused, the more their speech features exhibit differences with low arousal states.

As regards the analysis on Bipolar Data, even if the limited number of enrolled patients does not allow to generalize, the obtained results showed that some statistically significant differences can be observed at every different description level. Some feature trends were observed, but some of them were not always coherent among the enrolled patients or the investigated tasks. Some features trends might be patient or task specific. Notwithstanding, it is important to highlight that at higher description levels some features showed coherent trends. In fact, differently from the features investigated at lower description levels, that showed some patient-specific trends, at higher levels some of the inter-state analysis highlighted coherent feature trends among the enrolled patients. These results could mean that a higher level of description might be needed to overcome the problem of high vocal variability. For this purpose, some higher levels of description could still be taken into account. For instance, the information gathered by a semantic analysis of the speech, together with the approaches here investigated, could lead to obtain an interesting and deeper knowledge of these phenomena.

Contents

[1] Part of this Chapter has been already published in [1–5].
[2] Part of this Chapter has been already published in [1–5].

[3]Part of this Chapter has been already published in [1–5].

List of Figures

List of Tables

Chapter 1

Fundamental of voice studies

1.1 The Process of Speech Production and Perception in Human Beings

The fundamental purpose, behind the production of a speech, is communicating, transmitting messages. In Figure 1.1 a simple diagram representing the process of speech production and the process of speech perception is reported. Speech-generation process begins with the talker's action of formulating, in his/her mind, a message that he/she wants to transmit to the listener by means of his/her speech. A mechanical equivalent of this process, consisting in the formulation of a message, would be the creation of a linguistic message that is expression of the thought message in words [14]. Then, in the next step the message is converted into an oral code. Such a step can be represented as a conversion of the linguistic code of the thought message into a set of sounds (phoneme sequence) that combine into words. In this step, the prosodic markers associated with the sounds and providing information about duration of sounds, loudness of sounds, and pitch accent, have to be produced together with the phonemes. At this point, since the oral code is ready to be emitted, the talker has to activate a series of muscles to allow vocal folds vibrating, when appropriate, and to shape the vocal tract to generate the suitable sequence of speech sounds, producing as final output an acoustic signal. The activated muscles have to simultaneously control articulatory motion, such as lips, jaw, tongue and velum.

As soon as the speech signal is produced, emitted and propagated to the listener, the speech-perception process starts. First of all, the acoustic signal is processed by the listener along the basilar membrane in the inner ear, which is able to implement a running spectrum analysis of the received signal. Then the spectral decomposition of the signal at the output of the basilar membrane is transformed into activity signals on

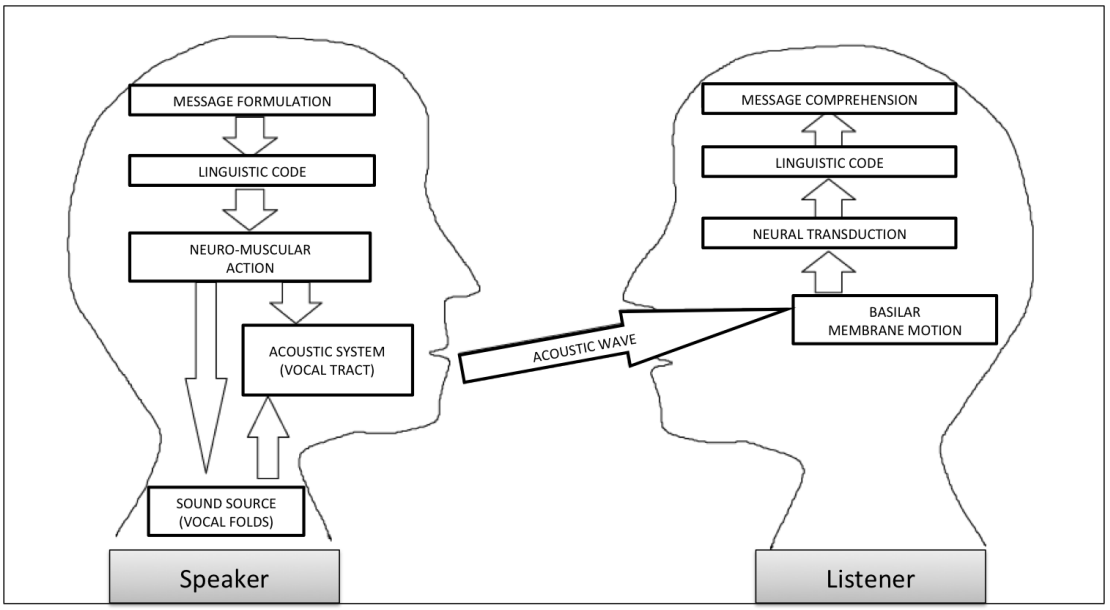

FIGURE 1.1: Mechanisms involved in the production and the perception of speech.

the auditory nerve. Such conversion is called neural transduction and corresponds to a feature extraction process. At this point, a not well understood process takes place. In fact, the neural activity along the auditory nerve is converted into a linguistic code at the higher centres of processing within the brain, and finally the message is understood.

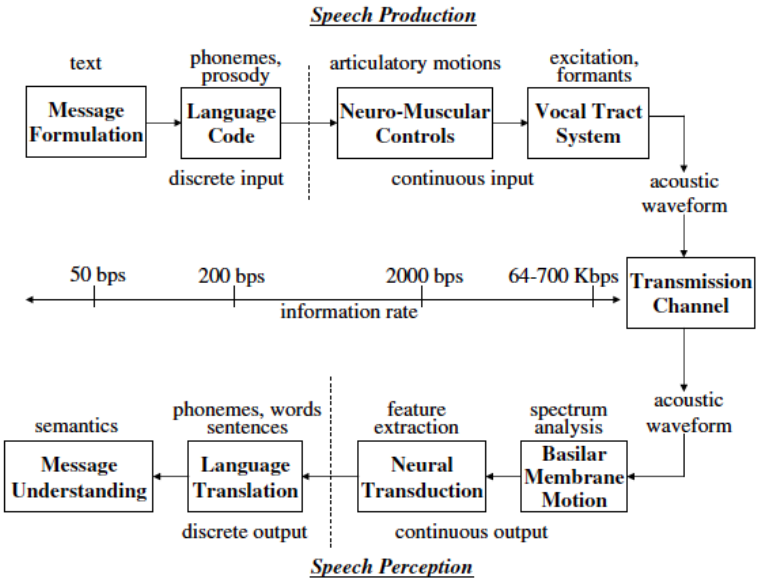

FIGURE 1.2: The Speech Chain: message, speech signal, and understanding [6].

A possible different way to view the speech-production/ speech-perception process is the one reported in Figure 1.2 [6]. In this figure all the steps previously described are reported along a line corresponding to the basic information rate of the signal at various stages of the process. Regarding the speech production, it is possible to estimate the

rate of information [6] assuming that the written message is formed of about 32 symbols, i.e. about 32 letters. In fact, in English there are 26 letters, but considering simple punctuation the number of symbols can be considered as equal to $32{=}2^5$. Moreover, on average the speaking rate is about, at a first approximation, 10 *symbols per second*. Therefore, if independent letters are assumed as a simple approximation, the base information rate of the written message can be considered to be equal to about 50 *bps*, i.e. 5 *bits* per symbol times 10 *symbols per second*. Then, at the following step, the text representation is converted into phonemes and prosody markers, e.g. pitch and stress. Here the information rate increases by a factor of 4 to about 200 *bps*.

In the first two stages of the speech chain, the representation of the information, that the speaker wants to convey, are discrete. Hence an estimation of the rate of information flow can be readily reached providing that some simple assumptions are made. In the next steps, where speech is produced, the representation becomes continuous, since articulatory motions produce the signal. The variation of the acoustic waveform appears to be faster than the articulatory movement. Some estimation of the bandwidth and required accuracy suggested that the data rate of the sampled articulatory control signal is about 2000 *bps* [11]. Hence, the spoken message requires a much higher rate than the rate that was estimated for the transmission of the written message. Such a high data rate is needed to represent the continuously varying signal. It is important to notice that, here, the term data rate refers to discrete representations, while the term information rate refers to the message. Finally, the data rate of the sampled speech waveform can be anywhere from 64000 to more than 700000 *bps*. In fact, for instance, in telephony it is required a "telephone quality" whose bandwidth is $0{-}4$ kHz, with a sampling rate of 8 kHz, and a resolution of 8 *bits* on a log scale. Such a quality results in a bit rate equal to 64000 *bps*. Such a representation of the speech signal is highly intelligible, but most listeners will be able to detect some differences with the original speech. Some other format of audio signals can be of higher quality: "CD quality" with a sampling rate equal to 44100 Hz and a bit resolution of 16 *bit* has a data rate of 705600 *bps*. Lately "high fidelity pure audio" systems are able to acquire and reproduce an audio signal with a resolution of 24 *bits* and sampled at a frequency equal to 192 kHz.

Moving through the speech chain, from the message to speech waveform, it is possible to observe a encoding of the message into an acoustic wave that can be propagated toward the listener. At this point the acoustic wave will be robustly decoded by the hearing mechanism of the listener. The above shows that the data rate can increase by a factor of 10000 passing from the linguistic code to the sampled speech waveform. Such increase in data rate can be partially explained taking into account some important characteristics of the talker, such as emotional state, speech mannerisms, accent, etc..

On the other hand, regarding speech perception, the model is formed by a series of steps from capturing speech at the ear to understanding the message encoded. Equally they can also be represented in terms of data or information rates. Hence, the continuous data rate at the basilar membrane is about $30000-50000$ *bps*, while at the neural transduction stage it is about 2000 *bps*. At the end, at the higher-processing-level within the brain, the neural signals is converted into a discrete representation, which can be decoded into a low-bit-rate message.

1.2 Voice production

In order to well understand the voice signal, it is very important to have clearly in mind how breathing works, since the source providing the raw power for speech and voice production is the *lungs*. Surely the main purpose of the breathing system is not speaking, but providing periodically the necessary amount of oxygen by means of the gas exchanges happening into the lungs, the sense of smell, and finally the regulation of temperature by means of the air flowing [7]. In this frame, the most important actors on the scene are the lungs, the muscle of the chest wall and the muscle of the abdomen. The lungs (Figure 1.3) are a complex network whose wide surface is the place where the gas exchange occurs. They are supported by the *diaphragm* (Figure 1.3) and enclosed by twelve pairs of *ribs* since they do not have any muscles of their own (Figure 1.3). The link between chest wall and lungs is made by the pleurae (Figure 1.3). In fact two layers, *visceral pleura* next to the lungs and *parietal pleura* next to the chest wall, are separated by a small layer of fluid. Such a structure allows either lungs and chest wall to slide against each other, but avoids that they separate. Therefore when the ribcage modifies its volume, the lungs are compressed or stretched accordingly.

All the ribs but two are connected to the sternum to form a closed cage. The other two, that are the lowest ones, are called free ribs because they are not attached as the others. Such a structure is flexible due to the cartilaginous attachments that allow the ribs to move. Each ribs is connected to other two: the one above and the one below. Such a connection is made by the *intercostal muscles*. The *external intercostal* muscles run from the bottom of one rib to the top of the next lower rib. The contraction of such muscles is responsible of the elevation of the ribs that expand the chest cavity. On the other hand, the *internal intercostal* muscles run from the bottom of each rib to the top of the next higher rib. The contraction of this kind of muscle results in a reduction of the size of the chest cavity. In Figure 1.4 intercostal muscles are shown.

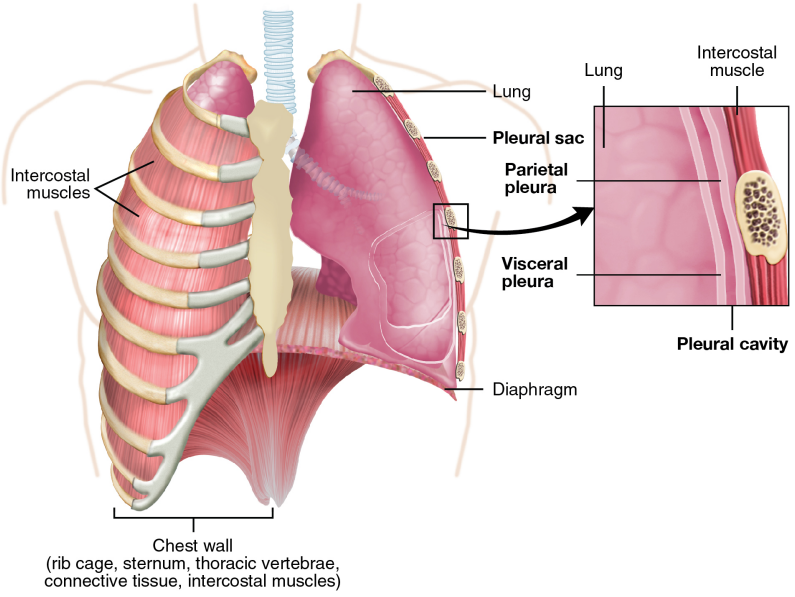

FIGURE 1.3: Lungs, pleurae, ribs, diaphragm, and intercostal muscles.

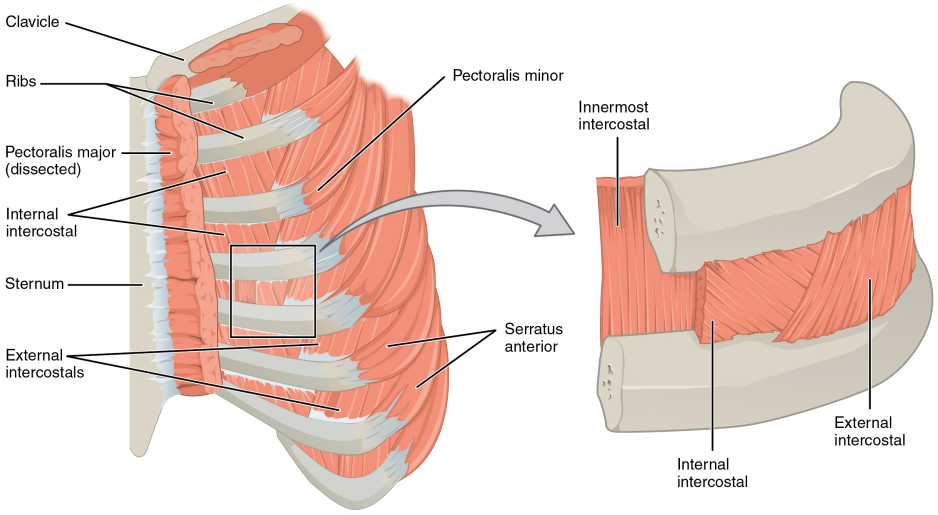

FIGURE 1.4: The ribs and intercostal muscles.

1.2.1 Breathing: in quiet condition and during speech

The air in the lungs, like any gas, is intrinsically elastic. The relation between pressure and volume at a constant temperature is determined according to Boyle's Law as inversely proportional. Therefore if the lungs expand, the pressure decreases. On the contrary, if the lungs reduce their volume, the pressure increases. Air always tries to reach an equilibrium point if this is possible, by flowing from a high pressure region to a low pressure one.

Within the lungs, pressure changes are created by means of the action of the respiratory system. The speaker can allow air flowing into or out of the lungs, by generating a

difference between the air pressure inside and outside the lungs. A lower inner pressure, with respect to the outer atmospheric air pressure, generates an air flow entering into the lungs. On the contrary, if the inner pressure is higher than the atmospheric one, air will flow out of the lungs. When respiration is quiet, the action of the respiratory muscles is able to increase the volume of the thoracic cavity, and therefore the lungs' one, and to decrease the air pressure in the lungs with respect to the atmospheric pressure. Such decreasing will generate an airflow into the lungs through the mouth and/or nose. Instead, when the chest cavity volume is decreased by the action of the respiratory system, the muscles compress the lungs. As a result the inner pressure will increase with respect to the outer one and the air will be forced out of the lungs. To sum up, the action of the muscles involved in respiration is not directly responsible of the air flowing into or out of the lungs. In fact, such muscular actions modify the thoracic cavity and the shape of the lungs. Such modifications generate a difference in air pressure that is able to activate in-and-out flow of air. On the other hand, when people inspire air, an active muscular contraction is always required. Such inspiration requires the contraction of external intercostal muscles. Their activity will expand the chest activity. Another muscle that may be involved is the diaphragm (Figure 1.3). The diaphragm is a muscle that, thanks to its dome-shape, separates the chest from the abdominal cavity. Its contraction performs a downward movement that is able to flatten out the dome, to push out the abdomen, to increase the size of the chest cavity, and therefore to create a suction which draws in the air by expanding the lungs.

On the contrary, when people exhale, both muscular action and passive forces can play a role. In fact, on one hand an active muscular contraction can reduce the size of the chest cavity by pulling the ribcage down. In this case the internal intercostals and the muscles of the back are the effectors of such active movement. In addition, by allowing the diaphragm to relax and return to its resting position, the same result can be reached. The reason is that the diaphragm can only contract downward, it cannot rise by itself, but the abdominal muscles are able to push the guts upward, pressing up on the diaphragm and thus inducing a decrease in the size of the chest cavity. On the other hand, the properties of the lung tissue can push air out of the lung without the application of any muscular forces. In fact, the naturally elastic property of lung tissue can contribute to pressing air out of the lungs. When stretched, this tissue will shrink back to its original shape. Such a property is called *elastic recoil*. In addition, the ribcage shows some elastic properties too. In fact the ribs' cartilaginous attachments allow the ribs to return to their resting configuration if no muscular action is opposed. In such a context, gravity also tends to pull the ribs down in absence of muscular contraction, while the contents of the abdomen might press upward on the diaphragm at the end of inhalation, shrinking the chest cavity. Such forces are dependent on the

quantity of air in the lungs. In fact, according to the elastic properties of the tissues, air, lung-stretching and passive recoil forces are directly proportional. In Figure 1.5 a scheme explaining the respiratory mechanism is shown.

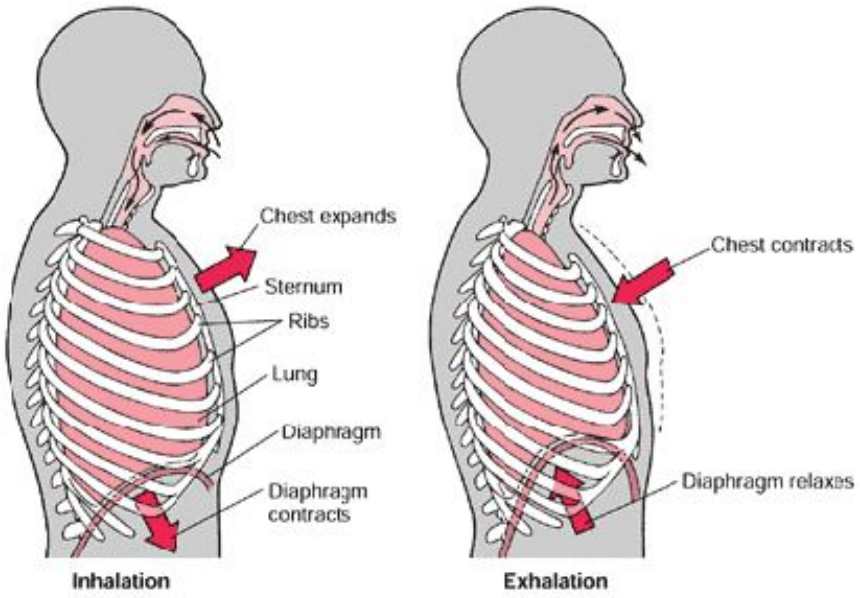

FIGURE 1.5: Scheme of muscular activation during respiration.

It is important to notice that, in absence of muscular effort, the quantity of air conveyed from the lungs to the vocal tract depends on the volume of air in the lungs [7]. The maximum recoil pressure is generated when the lungs are full of air. In fact, in such a case, the air pressure within the lungs is the highest possible and the lungs deliver the highest volume of flow. Alike, if the lungs are about in their resting state, the pressures are lower and the flow decreases.

People do not usually breath in a deep way during quiet respiration. Therefore elastic recoil, torque from the ribs, which rotate slightly on inspiration, and gravity are able to develop a pressure. Such a pressure is sufficient to induce an outward flow of air. Hence, expiration muscles are usually not necessary to breath out. Anyway, muscular activation is required to allow the thorax expanding for breathing in or breathing deeply out.

Breathing during speech is very different from quiet respiration. First of all, people when they are planning to speak, they usually breathe in more air than they do during quiet respiration [7]. This is also noticeable if relative durations on inhalation and exhalation are observed. When people breathe quietly, about 40% of a breath cycle is dedicated to inspiration and the remaining 60% of the cycle is dedicated to expiration. Instead, during speech, the inspiration phase can last only about 10% of a breath cycle. In fact,

exhalation takes longer to allow the production of on utterances with minimal interruption, and often it extends beyond the normal resting capacity of the lungs [20]. Often an expiratory muscular effort is needed in speech breathing. Such effort aims at producing enough airflow to complete an utterance. In fact, elastic recoil alone is not enough to induce an airflow able to perform a very long uninterrupted monologue. Moreover, a fairly constant relative pressures above and below the vocal chords is required by speech [21]. On the other hand, airflow and pressure during quiet respiration show a greater variability from the beginning to its end. Finally, the resistance to the flow above the vocal folds is modified continuously during speech. Such changes are originated by the movements of the tongue, lips, jaw and soft palate. Pressure will increase throughout the system when the airway is constricted or is blocked (for example, during the phonation of the sounds [t] or [s]). Instead, the opening of the vocal tract will decrease the pressure (phonation of a neutral vowel). For these reasons, maintaining a consistent pressure below the vocal folds is tricky. In conclusion, speech breathing is much more complicated than quiet respiration, since it is not just "breathe in and then relax" [7].

The respiratory system needs a dynamic balance between active muscular control and the forces due to the varying passive elastic recoil in order to maintain a relatively constant pressure below the vocal folds [22–24]. Notwithstanding, the pattern of muscular activity, used to control air pressure, is not well known. Usually, an inspiratory effort can be required to mitigate the excessive expiratory pressure, due to a high lung volume and hence high recoil forces, during exhalation, while, on the contrary, an expiratory effort can be required when low lung volumes generate little or no recoil forces. Anyway, some studies report that the continuous balancing of recoil forces requires the use of both inspiratory and expiratory muscles throughout a respiratory cycle [24–26]. These small muscular adjustments are aimed at providing a fast and continuous correction of sub-glottal pressure (the air pressure measured below the vocal folds). Such adjustments are provided to respond to the varying supra-glottal resistance during speech. In particular, during the expiratory phase of speech breathing, the abdominal muscles are active and help to control the sub-glottal pressure [25, 26]. The precise timing of these muscular activation is unknown, but they are thought to be speaker dependent.

This is a simplified and brief description of the respiratory activity during speech. In reality, speech is more complicated. Speakers, individually, prefer different muscular activity, and therefore, their muscular patterns can be significantly different. The greater variability in speaking was measured in the elderly [27]. The organization of breathing patterns for speech can differ in speakers. Many people organize the air intake in accordance with the length of the planned utterance [28, 29]. When people are asked to read out loud, they have different strategies. People usually vary, in a person-specific way, the lung volumes to perform louder utterances, inspirations at sentence and paragraph

boundaries, and length of utterance [30]. Some speakers can use their abdominal muscles for breathing, while others rely primarily on ribcage motions to vary the volume of the chest cavity. Such different behavior begin in infancy [31] and seem to persist during adulthood. Other factors influencing the pattern of muscular activation and thoracic wall movement are the speaker's age, body type and shape and finally the posture. For example, when the speaker is upright, gravity acts on the diaphragm and ribs. Respiratory activity can also influence the loudness. Moreover, a particular airflow can be requested to generate a particular sound: more airflow is required to produce [h] than to produce [t]. Lung volume can be influenced by linguistic factors, such as structural (clausal) boundary [28, 32]. Speaking seems admitting different breathing patterns. During listening to a conversational interaction, the duration of inspiratory acts were seen to be quite similar to the ones observed during speaking, while the breathing cycles of the partners became synchronous [33]. Several topics can be addressed in this research area. In fact, studies on the movements of the thoracic wall or of the abdomen could usefully be conducted jointly with other ones on the action potentials of respiratory muscles (for example, [34]).

1.3 Larynx and Phonation

When the respiratory system has generated a controlled airflow from the lungs, vocal folds can convert it to a sound. $Vocal\ folds$ are located within the $larynx$, also known as *voice box*. They can be referred to as *vocal cords*, but anyway, the term $folds$ should be preferred since it provides a better description of the involved structures. In fact, vocal folds are small folds of tissue, not strings, that oscillate alternatively to close and open the air-path. Such oscillations continuously interrupt the flow of air from the lungs and create air pressure changes that listeners can hear as a sound. The so described process of modulation of airflow, performed by the vocal folds, is called *phonation*. In addition, the sounds that are produced by such modulating actions of the vocal folds are known as *voiced*. In order to understand the complex process behind the voice production, it is important to know about the anatomy of the vocal tract and about the biomechanics of vocal fold vibration. Not all the sounds that one can produce can be defined as voiced. Sounds like [a], [z], and [m] are all produced by making vibrate the vocal fold, but other ones, for instance [h], and [s] are not produced with the vibrating vocal folds. Therefore they are known as *voiceless* or *unvoiced*. They are part of human speech. The vocal folds vibration, responsible of the production of voiced sounds, can be perceived touching with the fingers the neck close to the larynx.

1.3.1 The larynx

The transducer involved in voice production is located within the larynx. In fact, this is the place where the vocal folds are suspended by muscles, ligaments and membranes from the *hyoid bone* in the neck (see Figure 1.6).

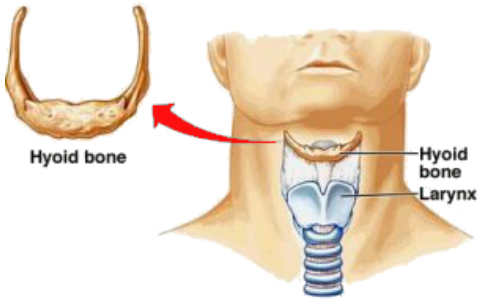

FIGURE 1.6: Location of the larynx and hyoid bone in the neck.

The larynx is fundamental in voice production, but its most important function is concerning the respiratory system. In fact the larynx completely seals the airway to protect both airway and lungs from foreign materials, particularly during swallowing. The larynx works also as a controlling valve of the quantity of airflow through the system. In fact, for example, the breathy voice produced during running, is caused by the larynx during increased rate of gas exchange during physical activity. Normal phonation can be produced during a consistently conditioned pace. Therefore, the vocal function of the larynx is simply supplementary to these older, more elemental, and biologically essential functions. Any structural change must maintain the fundamental ability to protect the airway, and hence the sound-producing potentialities of laryngeal structures has a limited extent. In fact, from a evolutionary perspective, the larynx is a rather conservative structure, and its anatomy is quite similar across all mammalian species [35]. A different structure is observable in birds. Birds have both a larynx and a separate organ, the syrinx, that is the dedicated organ to sound production. Since the syrinx is free of "multi-use constrains", its structure evolved considerably in different ways across birds species [35, 36].

A set of interconnected cartilages composes the larynx. They are placed in the airway below the pharynx and above the trachea (windpipe) in the neck. Its position in the neck is easily detectable, since if we try saying "ah", we can feel some vibration when placing our fingers on the neck. In fact, moving toward or away from the larynx, we should feel the vibration getting stronger or weaker. Moreover, during swallowing we can feel our fingers moving upward and downward.

The larynx is placed in the neck in suspension from the hyoid bone (Figures 1.6 and 1.7), placed just under and approximately parallel to the jaw. This bone serves a unique

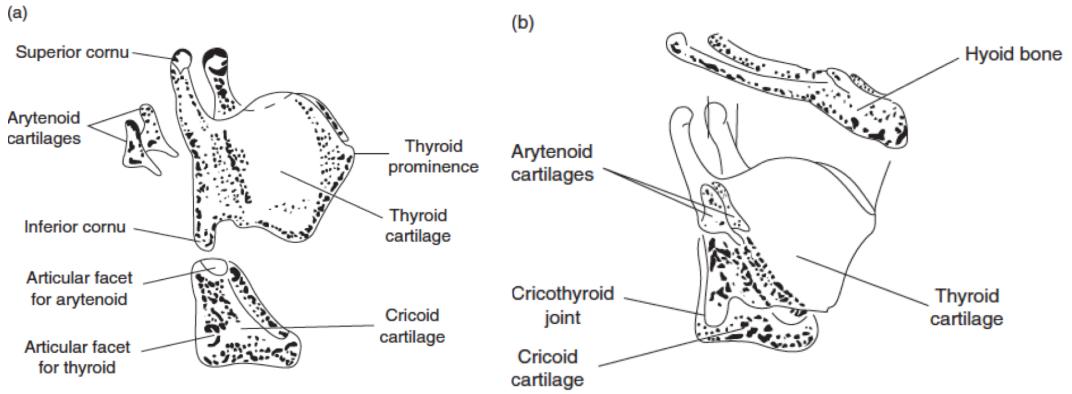

FIGURE 1.7: Disassembled (a) and reassembled (b) laryngeal cartilages, including the hyoid bone [7].

function, as no other one, since it does not connect or form a joint directly with any other bone or cartilage in the body. The hyoid bone is a point where more than twenty different muscles are attached. In addition, this bone can also protect the airway from injury.

The *thyroid cartilage* is the largest laryngeal cartilage (Figure 1.7). Its shape is something like a opened shield towards the back. On the back of the thyroid cartilage two sets of horns, or *cornua*, are placed. More specifically, two of them are placed on top (*superior*) and the other ones on the bottom (*inferior*). The superior cornua are connected to the hyoid bone by means of ligaments, while the inferior cornua are attached to the cricoid cartilage, as described subsequently. The union of the two sides (*laminae*) of the shield form an angle called the *thyroid prominence*, which is easily detectable in many men as "Adam's apple" (see Figure 1.8). Although it is less visible in women and children, it can still be located easily by touching that area of the neck. Usually the angle between the two laminae ranges from about 90° in men to about 120° in women [7]. This is the largest of the laryngeal cartilages, but it is still quite small. Size can vary considerably across persons, but the largest differences are between men and women. In males the average distance from the tips of the inferior cornua to the tips of the superior ones is 44 mm, while in females it is equal to 38 mm. In addition, the average anterior-posterior dimension is 37 mm in males, and 29 mm in females [37].

The *cricoid cartilage* is the second laryngeal cartilage. This is a signet-ring-shaped cartilage, whose broad part faces towards the rear. On average, the width at the back of the cricoid cartilage is 25 mm, while in front it is 8 mm [38]. Large differences can be detected across people [38]. The attachment of the thyroid cartilage, by means of the inferior cornua, to the cricoid cartilage is know as the *cricothyroid joint* (Figure 1.16). The larger posterior part of the cricoid cartilage is made to fit the back opening of the thyroid cartilage, while the front part of the cricoid cartilage fits just below the lower

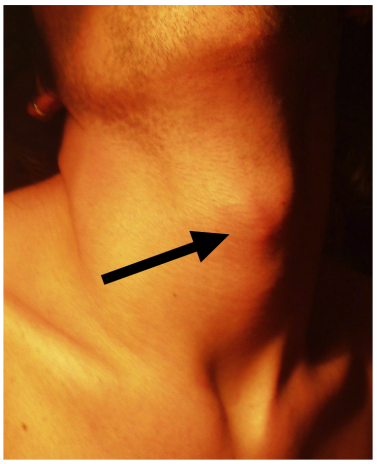

FIGURE 1.8: Position of "Adam's apple" in the neck.

part of the thyroid cartilage. The thyroid and cricoid cartilages can oscillate about the cricothyroid joint.

The final major components of the laryngeal framework are the two *arytenoid cartilages*. These paired cartilages are shaped like three-sided pyramids. On average, in men their height is about 18 mm, while in women it is 13 mm [37]. The arytenoid cartilages are placed on the top of the back of the cricoid cartilage. Each arytenoid cartilage forms with the cricoid cartilage a joint that is known as the *cricoarytenoid joint*. These joints are characterized by an extreme flexibility, since the arytenoids can move in several directions. Such a motion is usually defined as "rocking and gliding". Two small bumps (*processes*) project from the base of each arytenoid. The *muscular processes* are located at the back of each arytenoid cartilage, and at the front of the thyroid cartilage.

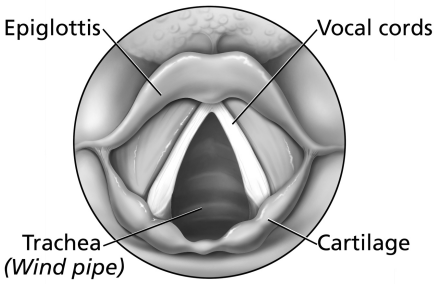

FIGURE 1.9: Location of the vocal folds in the larynx. View from above.

The vocal folds are placed across the airway, running from just below the thyroid notch to the arytenoid cartilages, where they attach to the vocal processes (Figure 1.9). The vocal folds are tiny. On average, their length is from about 17 mm to about 24 mm in males, while in females their length is about $13-17$ mm [39, 40]. Anyway they are characterized by a certain degree of flexibility, since they can stretch by about $3-4$ mm. Usually the space between the two vocal folds is know as the *glottis*, while *supraglottal*

is called the space above them, and *subglottal* the space below them (Figure 1.10). Around two-thirds of the glottal opening, usually referred as the *membranous glottis*, run between the vocal folds. The remaining one-third of the glottis runs between the two arytenoid cartilages, which project into the airway. *Cartilaginous glottis* is how this part is called. *Laryngeal ventricle*, or *ventricle of Morgagni* is the name of the space just above the folds, while above the ventricle the *ventricular folds* (also called the *false vocal folds*) are placed. *Conus elasticus* is the name of the space below the vocal folds.

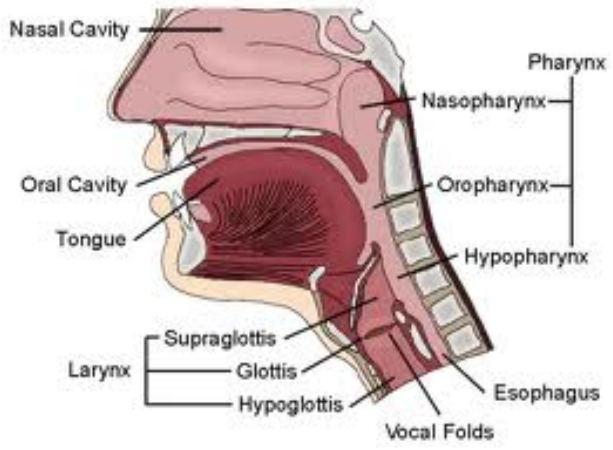

FIGURE 1.10: Supraglottal, glottis and subglottal spaces.

The structure of the vocal folds is quite complex and layered (Figure 1.11), and it is described by the *body−cover model* [41, 42]. The top layer of the vocal folds, i.e. the cover, is formed by two different parts: the *ephitelium* and the *lamina propria*. The ephitelium is a thin, stiff layer (about 0.05 *mm* thick) whose aim is protecting the folds from impact stresses and friction. The lamina propria, is about 1.5−2 *mm* thick and it is characterized by a layered structure. Each layer has small difference in its composition. These differences results in a distinctive vibratory behaviour. The *mucosa*, i.e. the topmost layer of the lamina propria, is very squishy and stretchable in all directions. On the contrary, the middle layer can be only stretched along the anterior-posterior axis, while the deepest layer resists stretching altogether. The body of the vocal folds is located below the cover, and it consists of the *thyroarytenoid muscles*, also known as *vocalis* muscle. Technically, only the median part of the thyroarytenoid muscle is referred as "vocalis", but the term are often used interchangeably (anatomic details of this muscle are discussed in [43]). This muscle is the bulk of the vocal folds. The different mechanical properties of these layers of tissue entail that the body of the vocal folds can be stiffened, while, on the contrary, the external cover remains loose and is free to move around the body. This mechanical property is fundamental for phonation.

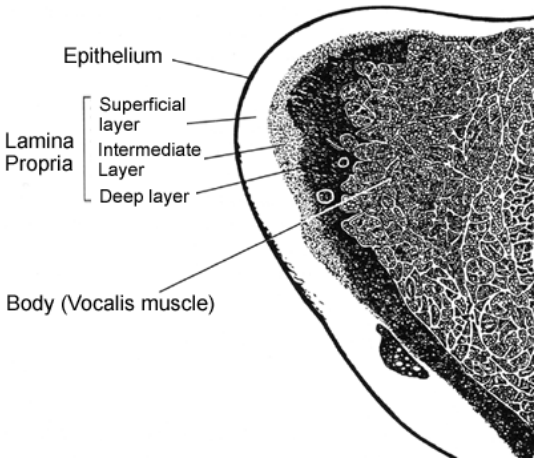

FIGURE 1.11: Structure of vocal folds: body-cover model [8].

The laryngeal muscles can be divided into two groups: the *intrinsic laryngeal muscles* and the *extrinsic laryngeal muscles*.The role of the extrinsic muscles is connecting the larynx to other parts of the body, stabilizing its position and moving it up and down. Instead, the role of the intrinsic muscles is connecting the different cartilages and changing their relative positions. This latter kind of muscles is formed by some of the smallest and fastest muscles in the body. Moreover, these muscles provide very quick and precise control of position of the laryngeal cartilages. As a result, during breathing the vocal folds are able to produce complete glottal opening, or complete glottal closure to support lifting or protect the airway. In addition, these muscles may favour vocal fold vibration to produce sound, or narrow the glottal opening to produce whisper.

In Figure 1.12 the movements of the vocal folds are shown during breathing. Instead, during phonation the folds must be positioned at midline and they must move apart for the next breath. Usually these two different functions are reflected in two different sub-groups of intrinsic muscles. The laryngeal *adductors* aim at bringing the folds together, while the *abductors*' aim is pulling them apart. The position of these muscles are displayed in Figures 1.13, 1.12, 1.14, 1.15, 1.16 [39].

The interarytenoid and the lateral cricoarytenoid muscles are the two primary laryngeal adductors. Usually, muscles are named taking into account the structures they connect. Therefore, the lateral cricoarytenoid is the muscle running between the muscular process of each arytenoid and the side of the cricoid cartilage, while the interarytenoid muscle is the one running between the muscular process of the two arytenoid cartilages (Figure 1.13).

The backs of the arytenoids slide towards each other when contracting the interarytenoid muscle. As a result, a gap may appear in the membranous glottis. During the contraction of the interarytenoid muscle, the position of the membranous vocal folds change just

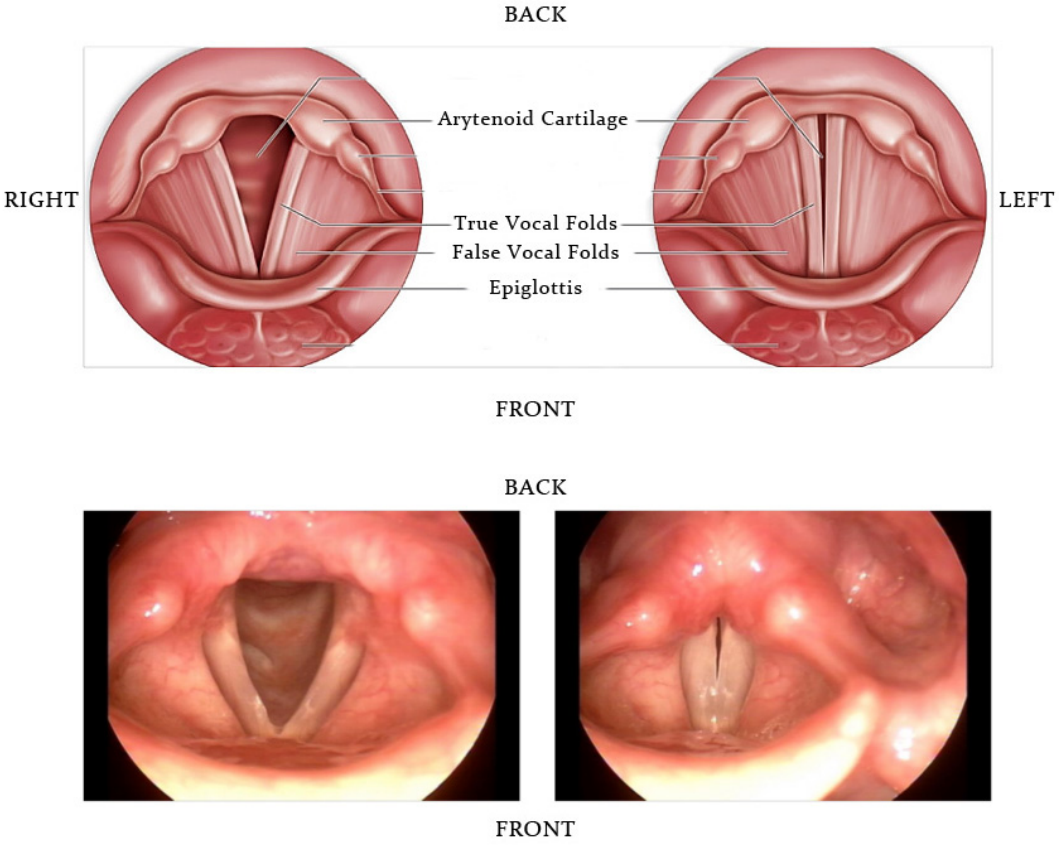

FIGURE 1.12: Vocal fold motion during breathing.

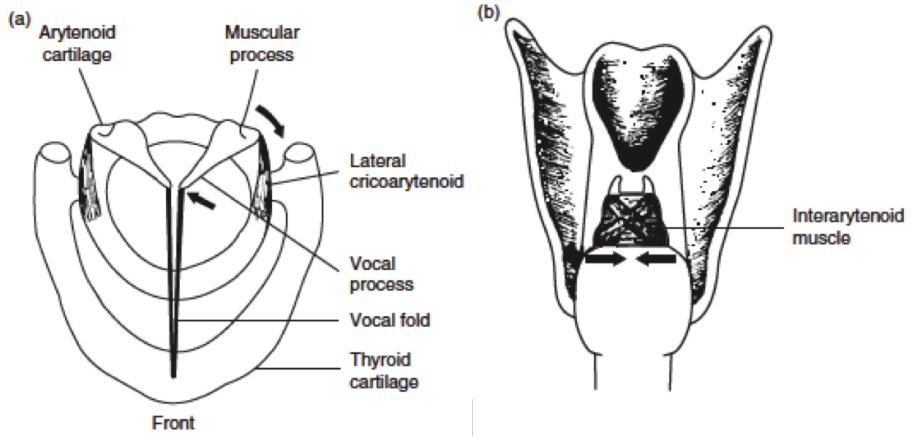

FIGURE 1.13: The lateral cricoarytenoid muscle (a), the interarytenoid (b) muscles and the muscular actions being able to produce the so calling "rocking and gliding" motion of the arytenoid cartilages [7].

a little. Moreover, when the interarytenoid and lateral cricoarytenoid are contracted together, the posterior part of the glottis is closed by their joint action, but a medial gap between the vocal folds may remain.

The thyroarytenoid muscle is located between the vocal process of each arytenoid and

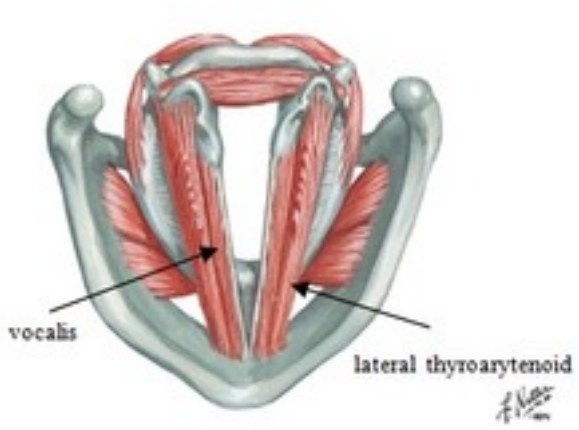

FIGURE 1.14: The thyroarytenoid muscle [9].

the front of the thyroid cartilage (Figure 1.14).The deepest layer and the main body of the vocal folds are formed by this muscle (Figure 1.9). Within the muscle, some separate functional units are distinguished by some authors. Anyway, the complex structure and functions of these units are not completely understood [39, 43]. The contraction of this muscle generates a tension in the body of the vocal folds and a slight bunching of them. Such an action is demonstrated for each fold separately and for both sides together. The protuberant occurring with the contraction of the thyroarytenoid muscle also closes the central portion of the glottis, so that the thyroarytenoid is sometimes grouped with the laryngeal adductors. The contraction manifests also a secondary effect. In fact, the stiffness of the cover relative to the body is reduced. During phonation, a critical phenomenon is the control of the relative stiffness of the different layers of the vocal folds.

Usually, the joint actions of the cricoarytenoid, interarytenoid and thyroarytenoid muscles is required to achieve a complete closure of the glottis. This can be demonstrated by performing a stimulation of the branch of the recurrent laryngeal nerve that innervates these muscles, before the branchings to the individual muscles.

The posterior cricoarytenoid muscle is the only laryngeal abductor (Figure 1.15). It is located between the muscular process of each arytenoid and the back of the cricoid cartilage. The action of the posterior cricoarytenoid muscle pulls the back of the arytenoid cartilages somewhat medially and down, resulting in a vocal process swung upward and away from the midline. At the end of such action, the glottis is widely open, as for a deep or sudden breath. An opposite action of the cricoarytenoid muscle may also be performed by anchoring the arytenoids during phonation by means of the action of the posterior cricoarytenoid.

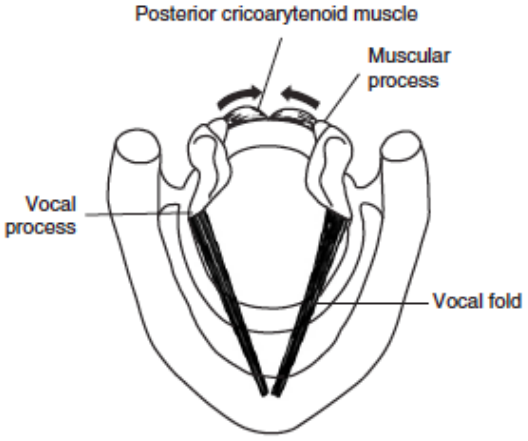

FIGURE 1.15: The posterior cricoarytenoid muscle and its actions (top view) [7].

One remaining intrinsic laryngeal muscle must be described, the *cricothyroid muscle*. This muscle can not be considered neither an abductor nor an adductor. It connects the thyroid and cricoid cartilages (Figure 1.16). Its contraction tilts the front of the two cartilages toward each other resulting in an increase of the distance between the front of the thyroid cartilage and the arytenoids, and in a stretching of the vocal folds. The nerve innervating the cricothyroid muscle is the *superior laryngeal nerve* (Figure 1.16). This is the nerve that is responsible for controlling F0. The *recurrent laryngeal nerve* innervates all the other intrinsic laryngeal muscles. It is important to note that both the recurrent and superior laryngeal nerves originate from the *vagus nerve* (cranial nerve X), which is the controller nerve of all the intrinsic laryngeal muscles [44]

In conclusion, the larynx is a connection of the following cartilages: the thyroid, the cricoid and the arytenoid cartilages. The larynx is suspended from the hyoid bone and is placed on the top of the trachea and in front of the esophagus. The relative position of these cartilages is controlled by the intrinsic laryngeal muscles. In fact, the vocal folds are closed by the adductors for phonation or to protect the airways, while the abductors opens them for breathing. The thyroarytenoid muscle is the most important intrinsic laryngeal muscle. It is used to control loudness and to change vocal quality, in the narrow sense. On the other hand, the lateral cricoarytenoid and the interarytenoid are the main adductor muscles. The former brings the body of the vocal folds together. It acts by pulling the back of the arytenoid sideways and down, and it brings together the edges of the arytenoids (where the vocal folds are attached). The gap between the backs of the arytenoid cartilages is closed by the interarytenoid muscle. The vocal folds are stretched by the cricoarytenoid muscle by tilting the thyroid and cricoid cartilages. The posterior cricoarytenoid is the only abductor muscle, and it opens the airway by pulling the back of the arytenoid cartilages down and toward the midline, resulting in a separation of the front edges.

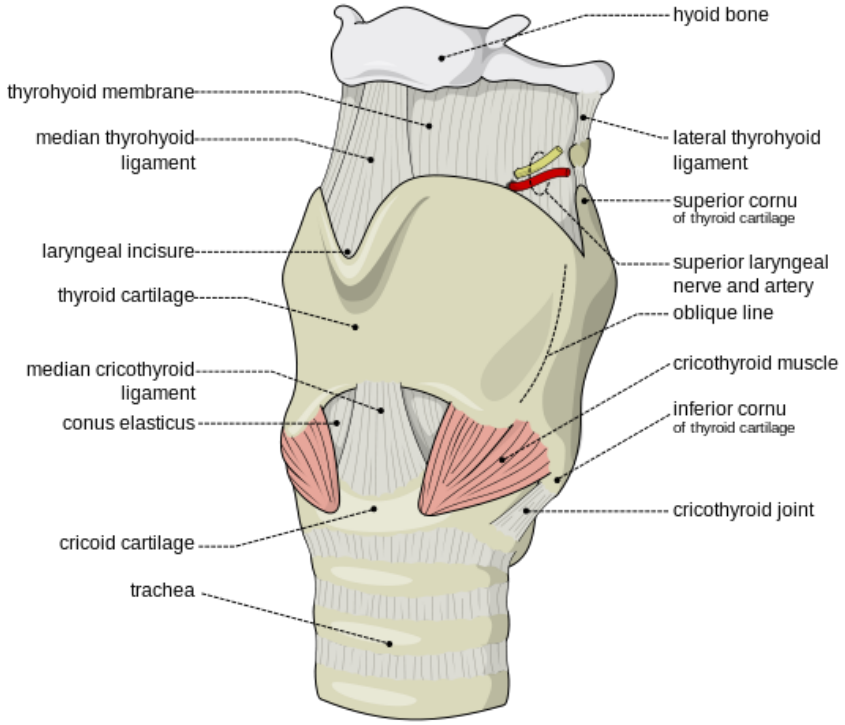

FIGURE 1.16: The cricothyroid muscle.

1.3.2 Mechanism of the vibration of the vocal folds

The contraction of the intrinsic laryngeal muscles acts to bring the vocal folds together at the middle of the glottis. In addition, they can stiffen the folds, but such actions can not produce by themselves vibration or sound. On the contrary, the sustained tissue oscillations, producing the voice, is generated by a combination of tissue elasticity and aerodynamic forces. In fact, when the vocal folds oscillate, they are able to periodically interrupt the flow in the airway from the lungs. The result of such interruptions is the creation of changes in the air pressure, changes generated by opening and closure rather than by pressing on air particles as vibrating strings do. This operative difference between vibrating strings and vocals folds is the reason why the term "vocal cords" is dispreferred, since it refers to an incorrect vibration type. The *myoelastic aerodynamic theory of vocal folds vibration* [45] describes these biomechanical and aerodynamic forces.

The first step to start vocal folds vibrating has to be performed by the action of the adductor muscles. In fact, first of all, these muscles have to bring the vocal folds together, so that they are closed or nearly so, but not held together too tightly. Successively, excess air pressure below the closed vocal folds is generated by the action of the respiratory system. On average, a subglottal pressure of about $3-5\ cm\ H_20$ is required to initiate vibration, but a lower pressure will be needed once vibration has started [42, 46]. In

fact, when the pressure is able to overcome the stiffness of the folds and the inertia of the air column above the folds, the vocal folds are blown open from below. Otherwise, if the vocal folds are too stiff (held together too tightly), or if the pressure below the folds is too low, they will not open. Another possibility is when the vocal folds are not stiff enough. In such a case they will blow open, and remain open. Therefore the balance between stiffness and subglottal air pressure has to be the proper one.

Two factors act to close the glottis after the air flows through the opening glottis. The tissue elasticity is the first one. The air pressure blows apart laterally the vocal folds from below, but their elasticity acts to let them naturally regain their original position at the midline as air pressure decreases with the free flowing of air through the glottis. Moreover, another factor aiming at the closing of the vocal folds is the contribution given by the aerodynamic forces. Such a phenomenon can be described according to Bernoulli, who, in the principle known with his name, stated that when air particle velocity increases, the pressure must decrease, as long as total energy remains equal. Hence, according to Bernoulli's principle, this is the reason of the reduction in pressure away from the midline which helps the vocal folds closing by sucking them back toward the midline. A further aeroacoustic contribution, the second one, to glottal closure occurs when vortices are formed in the airflow at the exit of the glottis. An additional negative pressure between the vocal folds is created by the vortices along the superior medial surface of the vocal folds. Such negative pressure contributes to quickly closing the folds and leads to an increase in the high-frequency energy in the voice source [47–49]. As soon as the vocal folds are closed, pressure starts again increasing below them until they are blown open and the whole cycle repeats once again. A $self-sustaining$ *oscillation* of the vocal folds is thus the result of these steps.

Vocal folds do not vibrate as a single unit moving rigidly back and forth across the glottis, but rather in a complicated way. The pressure below opens the folds from the bottom to the top, while the bottom edges close earlier than the top edges. In Figure 1.17 a single cycle of vocal fold vibration is shown, as viewed from the front of the neck. In subfigure 1, the vocal folds are closed and the air pressure below them is increasing. When pressure has a high enough value, it gradually begins opening the vocal folds from below, so that the edges start separating from the bottom (subfigures 2 and 3). In subfigure 4 the top edges of the vocal folds are separated and the air begins flowing through the glottis. At this point, Bernoulli forces, and vortices-related forces immediately begin acting to draw them back together. In subfigure 7, a lateral movement is still performed by the upper edges of the vocal folds, but the position of the bottom margins is near the midline again. The lower edges are again positioned in the midline in subfigure 8, and the glottis is closed. At this time, the air stops passing through the glottis, and the pressure starts increasing again below the folds, notwithstanding the upper edges of the vocal folds are

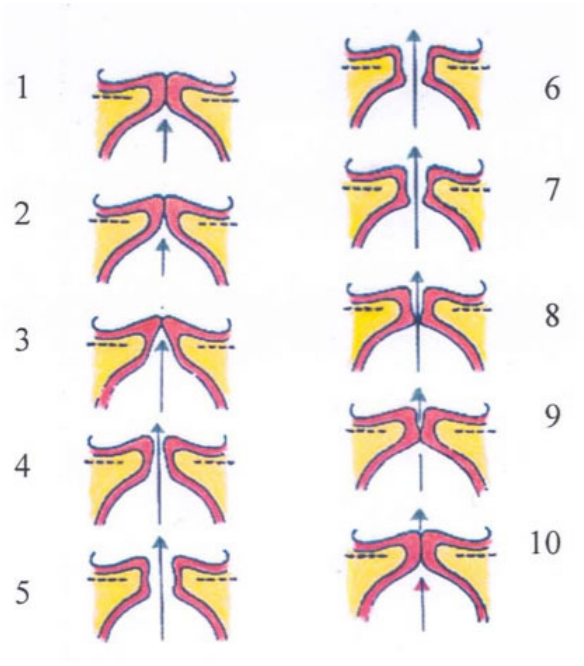

FIGURE 1.17: The movements of the vocal folds across the airway for a single cycle of phonation. [10].

still moving laterally, or may have only begun to return back to midline. Finally, in subfigure 10, the vocal folds have almost ceased to move laterally at the upper margin and the lower edges are beginning to separate again, giving a new start to this cycle [50].

The reason of such a way of vibrating of the vocal folds is due to their complex layered structure. The layer of the vocal folds that is primarily involved in the vibratory movement is the cover, which rides over the stiffer body. In fact, the difference in stiffness between the body and the cover controls the vibration. In particular, when there is no difference in stiffness (this can be originated by the slackness in the thyroarytenoid, due to paralysis, or stiffening of the mucosa, due to a respiratory infection), the vocal folds can not perform any vibration in a correct way. The relative stiffness of the cover and body of the vocal folds can be changed by adjusting the laryngeal muscles, resulting therefore in a change of the rate and pattern of vibration. The movement of the outer layer of the vocal folds, i.e. the cover, over the body layers is usually known as *mucosal wave*. Such a movement is easily detectable by means of stroboscopic images or high-speed films of the vibrating larynx. The description of the mucosal wave is usually focussed on its propagation along the superior surface of vocal folds. Such propagation is a visible effect when imaging techniques are used, and its speed depends on its direction. In fact, the wave is quickly attenuated travelling outward away from the glottis. Anyway, the most relevant part of the wave is its passage along the medial surface of the vocal folds, since the glottal opening and closing occur there. As it is possible to

notice in Figure 1.17, the wave begins there, and moves for a significant distance along the medial surface before "breaking" across the top of the vocal folds. The greatest amplitude of the wave is detectable along the medial surface of the folds, because of the interaction with the opposite vocal fold is during collision when the glottis is closing. It is important to highlight that most of the motion is in the mucosa, since the thyroarytenoid moves very little during vocal fold vibration. Such movements, occurring along the medial surface of the vocal folds, are the modulating factors of the airflow, and thus the phenomena in the production of acoustic waves, perceived as voice [50].

The production of sound is caused by laryngeal vibrations as described in the following. As soon as the vocal folds are opening, air is able to rush through the glottis and encounters the column of air that is above the folds in the vocal tract. The upward pressure of the flowing air against the column of air generates an increase in the supraglottal air pressure, i.e. a compression, and sets those air molecules in motion. The molecules, therefore, start moving, spreading out and rising up through the vocal tract. As the glottis closes, air molecules keep on moving above the glottis, thanks to their momentum, even if no further air is flowing through the glottis for the moment. This continuous movement generates a rarefaction, i.e. a decrease of the pressure above the vocal folds, since the molecules keep on moving rising up the vocal tract, but no further molecules is able to pass the closed vocal folds to maintain constant the pressure. So, the airflow from the lungs is modulated by the vibration of the vocal folds. Such a modulation creates a pattern of alternating compressions and rarefactions in the air flowing through the glottis which results, at the end, in the generation of a sound. A further sound source may be furnished by the vortexes in the airflow [47], but the mechanism behind this kind of sound generation is still not well understood, even if a growing attention is focussing regarding this topic [51–53].

The acoustic complexity of the voice signals is partially caused by the manner in which the vocal folds oscillate [7]. Since the period of a sound and the frequency are one the reciprocal of the other, events that happen with a short period, quickly, are necessary associated with high frequencies. Hence, the more quickly the vocal folds perform a complete oscillation opening-closing, the more high-frequency energy is produced by means of the brusque changes of the pressure. In fact, the closure of the glottis (Figure 1.17) normally happens rather abruptly. Notwithstanding the vocal folds may keep on moving after the stopping of the flow, these movement do not cause any further reduction in airflow since the closure is already complete. However, the vocal folds open more gradually (slower action), from bottom to top (Figure 1.17), resulting in lower frequency energies. Consequently, these differences indicate that *most of the acoustic energy in a normal voice is generated when the glottis closes, not when it opens* [7].

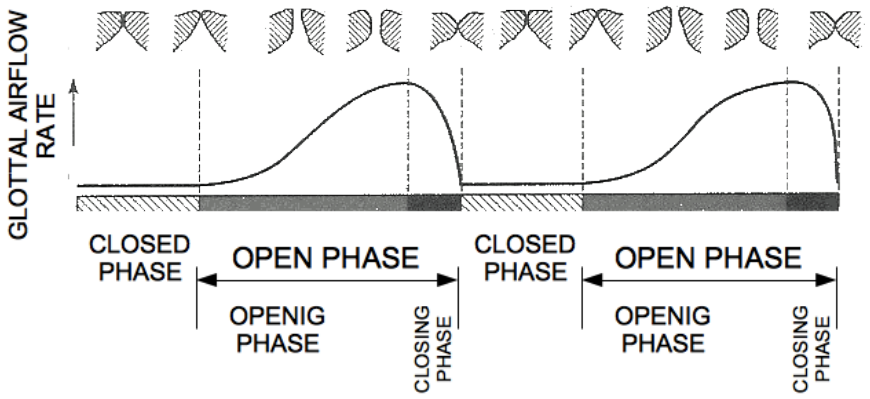

FIGURE 1.18: Schematic sequence for two vocal fold vibration cycle. Vocal fold vibration sequence is viewed as if viewed from the front and idealized glottal airflow waveform. Vocal fold opening, closing, open and closed phases are indicated.

In Figures 1.18 and 1.19 it is shown the airflow profile change through the glottis depending on the opening or closing of the vocal folds over time. In particular, in Figure 1.18, three key phases of the vibration cycle are detected: closed phase, opening phase and closing phase. The opening and closing phases are often considered as the 'open phase' of the glottis, because, during this phase, air flows. In addition, it is also important to notice that airflow is not necessarily null during the closed phase. In fact, there are vocal fold vibration patterns for which they do not come together over their whole length [54, 55]. In Figure 1.19 the airflow profile patterns for two different normal voices are reported. In (a) a pattern concerning a male voice is displayed, while in (c) a female one. For the male voice it is possible to recognize a quicker glottal closure (at the top of the Figure), while female voice is characterized by a slower closure (at the bottom of the Figure). It is also important to notice that the first waveform is asymmetrical: as the vocal folds open, the airflow increases relatively gradually, but it decreases suddenly as the folds close. Moreover, the folds also remain closed for a time interval, as marked by the lines near the value 0 between the pulses. On the contrary, the second waveform is pretty symmetrical, with its airflow increasing and decreasing lasting about the same time. In fact, the folds do not remain completely closed in this female airflow pattern, or may not achieve complete closure in every cycle. The subfigures on the right display the harmonic energy produced by the two voice sources (the *source spectra*). The harmonic differences in the two spectra depend on F0. Generally normal F0 is different in adults: females show a higher F0 value than males. Especially, in the Figure, the male voice is characterized by an F0 equal to about 115 Hz, while the female voice is characterized of an F0 equal to about 220 Hz. Since vocal fold vibration can be defined approximately periodic for both these voices, the frequency of each harmonic in the voice source is a whole-number multiple of the fundamental frequency. It is important to note that, in a

normal source spectrum, energy decreases approximately as a straight line, at a rate of about 6 dB/octave. Energy decrease in female voice is more quick in the lower part of the Figure. Voice can present strong and quick vocal fold closure and their related high-frequency harmonics. This kind of voice are usually described as "bright-sounding". On the contrary, voice can also present gradual or incomplete closure of the vocal folds. This kind of voice are usually characterized by having most of their energy at or near the fundamental frequency, and are described as "dull" or "weak".

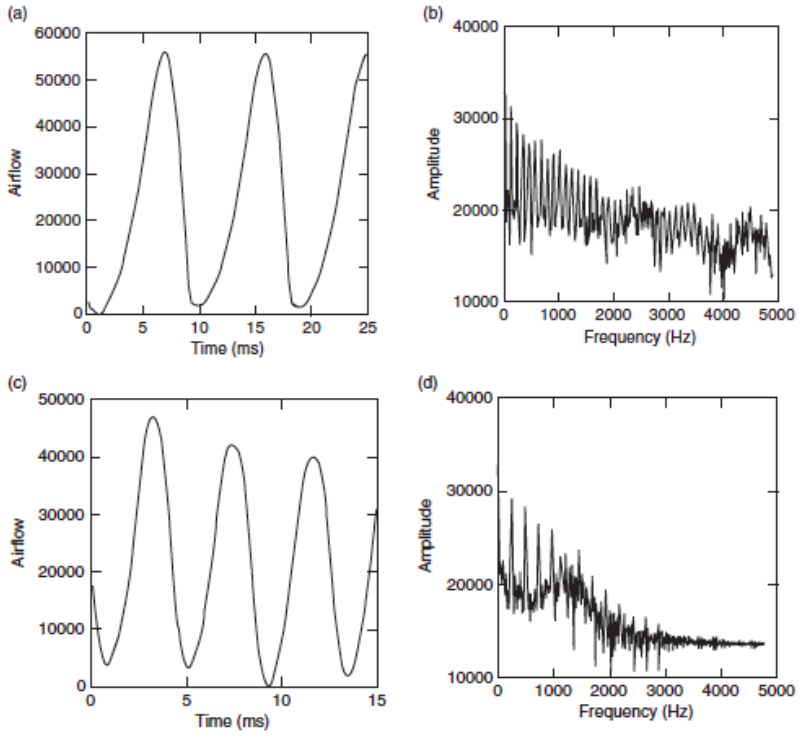

FIGURE 1.19: Pressure patterns of airflow during the passage through the glottis as the vocal folds open and close. Two different, but typical voices are considered: (a) and (b) take into account a male voice, while (c) and (d) a female one. In (a) the airflow pattern over time for a normal male voice is reported, and in (b) its spectrum is displayed. In (c) the airflow pattern over time for a normal female voice is reported and in (d) its spectrum is displayed. Arbitrary units are used in the y axes [7].

1.3.3 An introduction to a mechanical model of vocal tract

A way of representing the physiological mechanism of producing speech can be observed in the model in Figure 1.20 [11]. The lungs and the associated muscular action can be represented as the source of the airflow, shown as a piston pushing up within a cylinder. The air is pushed out of the lungs by the muscular force and through the bronchi and trachea. The vocal folds tension jointly with the air flow allow the vocal folds vibrating. Such vibration is the origin of the voiced speech sounds. The air is also able to produce a sound by passing through a constriction in the vocal tract. This constriction favours

turbulent airflow, that is able to produce unvoiced sounds. Moreover, the air pressure may increase behind a point of total closure within the vocal tract, and when the folds open, the sudden and abrupt release of the pressure is responsible of the generation of a brief transient sound.

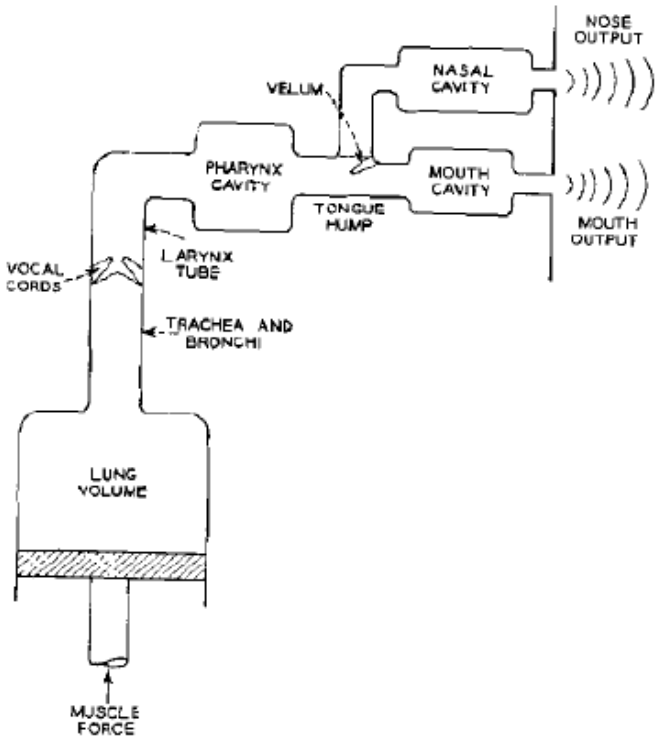

FIGURE 1.20: A mechanical model of vocal tract [11].

Since the speech usually is a sequence of sounds, the state of the vocal folds, their position, shape, and the size of the various articulators change over time to produce the sounds that have been programmed by the speaker. Such phenomena are addressed in the following section.

1.4 The Supraglottal Vocal Tract and Resonance

The supraglottal vocal tract (Figure 1.21), like the rest constituting the anatomy of speech and voice, is not designed for communicating. Lips, tongue, teeth and jaw are designed for feeding (chewing and swallowing food). Similarly, the nasal cavity has been mainly developed for smelling and for heating, humidifying and filtering air as it breathed in. Tongue, lips and jaw are able to move, modify and alter finely the shape of the oral cavity. For instance, the vocal tract can be made longer by the protrusion of the lips. Air can be conveyed or not through the nasal cavity. By lowering the soft palate, air, and sound energy, pass from the lungs through the nasal cavity, while by

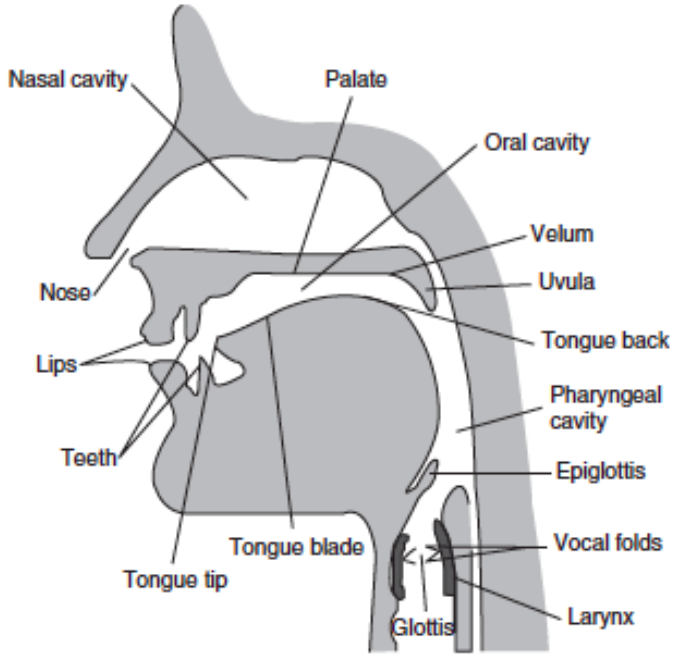

FIGURE 1.21: The Supraglottal Vocal Tract [7].

raising it, the nasal cavity is closed. The oral cavity can be enlarged by lowering the jaw, while the movement of tongue can change the shape of the whole cavity. All these modification of the shape of the vocal tract, via the motion of the articulators, change the sound that is emitted by the speaker.

The sound produced by the vibration of the vocal folds would remind a buzz more than a normal human voice if it was heard without the influence of the rest of the vocal tract. The buzzing sound is shaped by the acoustic effect of the vocal tract above the vocal folds. Hence, the vocal tract together with the vocal folds produces the sound that is heard, via *resonance*. In Figure 1.22 all the contribution are displayed. Lungs, by means of breathing provide the source of phonation. The laryinx modulates the airflow coming from the lungs. The vocal tract modifies the acoustic pressure, by means of its resonance.

Resonance is the amplification via constructive interference of waves. As a consequence, frequency components of the glottal source the frequency of which are near the resonance frequency are amplified. The spectral effects of vocal tract resonances are known as formants. Therefore a resonator can also be thought as filter. The frequencies close to its natural frequency pass, but the remaining are damped out.

The effect of a resonator on the spectrum of a acoustic source is reported in Figure 1.23. At the top the harmonics of the source are displayed (subfigure a). The resonator transfer function (b) modifies the source spectrum, while the result is displayed in subfigure c. A

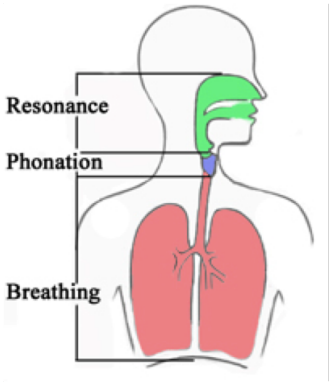

FIGURE 1.22: Scheme of the contribution of the different tracts in phonation.

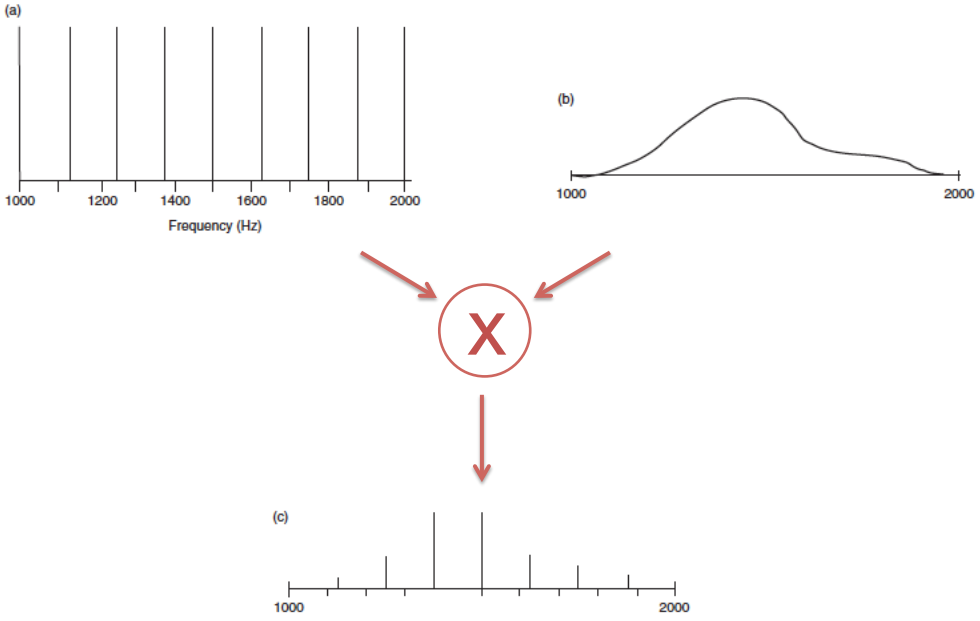

FIGURE 1.23: Vibrating source of acoustic energy and a resonator: interaction. The harmonics of the voice source are reported in (a), the resonator's frequency response in (b), and finally in (c) the result of exciting the resonator with the voicing source is displayed.

resonator has a *broad bandwidth* if it responds to a wide range of frequencies, otherwise it has a *narrow bandwidth* if it responds to a narrow range of frequencies.

Fant in 1960 [56] described how these principles represent the acoustic characteristics of speech sound in his *linear source−filter theory of speech production*. In this model the source is represented by pulsatile airflow or turbulent airflow. In Figure 1.24 (a) a typical voice source, comprising many harmonics, is reported. In Figure 1.24 (b) the resonance characteristics of the supraglottal vocal tract is shown. Its resonances depend on its length and shape. A vocal tract with a longer length is characterized by lower resonant frequencies with respect to a shorter one. The spectral effects of the resonances of the vocal tract are called *formants*, while their centre frequencies are

known as $formant\ frequencies$. In 1.24 (b) the formant frequencies are at about 500 Hz, 1500 Hz, and 2500 Hz and they correspond approximately to the vowel [ə] spoken by a male. The set of all the formants considered together is called the $vocal\ tract\ transfer\ function$. It describes how the source energy is transferred by the vocal tract to the outside. It describes the relationship between the acoustic input, provided by the vocal folds, and the output of the supralaryngeal vocal tract [7]. Many perceptually important details can be summarized just taking into account formants and their bandwidths.

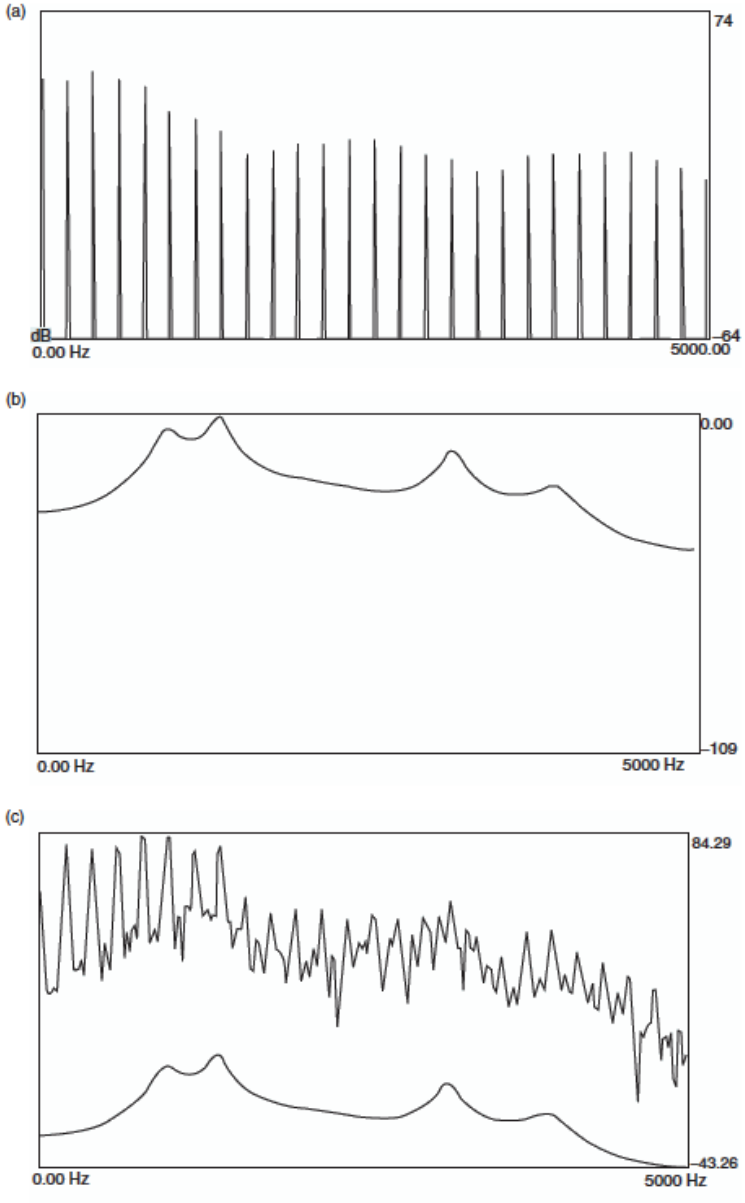

FIGURE 1.24: Source-filter theory of voice production. In (a) the spectrum of voice source is reported, in (b) the vocal tract transfer function, and finally in (c) the output voice spectrum, with the transfer function shown at the bottom of the frame, are displayed [7].

The pulsating airflow at the glottis sets the air in the vocal tract vibrating, acoustically

exciting the vocal tract resonances. Hence, the frequencies in the source waveform close to the resonance ones are amplified, while the other ones are damped. Every movement of the jaw, tongue, and/or soft palate or every tension of the pharyngeal muscles, and/or every shaping of the lips results in a modification of its resonant characteristics. Generally, differences in the first three formants ($F1$, $F2$, $F3$), within the same speaker, are associated with differences in vowels. On the contrary, consistent differences in formant frequencies and bandwidth found across speakers are associated with differences in personal voice quality.

A last component of the source-filter theory remains to be described, and this is the *radiation characteristic*. When the speech signal is radiated into space, it is emitted in all direction. The effect is the increase of the level of the higher-frequency part of the spectrum, equal to about 6 $dB/octave$. The result of the combination of the three components of the source-filter model, i.e. source, transfer function, and radiation characteristics, is reported at the bottom of Figure 1.24.

1.5 The Sound of Voice

The aerodynamic energy of the column of air is converted by the vibrating vocal folds into pulsatile airflow that is converted into acoustic energy. Finally, the acoustic energy, generated within the voice box, is shaped by the resonance properties of the supraglottal vocal tract to produce the sound that is heard.

Normally, three main perceptual characteristics are taken into account to describe the emitted sounds: pitch, loudness and quality. Every eventual modification of the mass, length and tension of the vocal folds can alter their vibratory behavior, and therefore modify the pitch, loudness and the quality of the voice. It is important to highlight that these three features are psychological characteristics, and hence they describe how physical signals are perceived.

1.5.1 Frequency and Pitch

Normally, the perceived *pitch* of a voice is mainly determined by the fundamental frequency (F0) of the source signal [57]. Listeners are very sensitive to changes in F0, and they are able to accurately detect changes of as little as 2% (2.4 Hz; [58]). From a mechanical point of view, F0 corresponds to the rate of vibration of the vocal folds. In Figure 1.25, on a piano keyboard some typical F0 ranges for different classes of voice, in speaking and singing, are displayed. On average, for males the F0 values is about 115 Hz. In females, the average F0 value is equal to about 220 Hz, while in children the

average F0 value is about 280 Hz [59]. In a singer, the F0 range can cover four or five octaves. F0 for a bass singer can be 80 Hz or lower, while for a soprano F0 can reach 1000 Hz. Perceived pitch does not depend only on F0, but it can also be influenced by the resonances of the vocal tract. In fact, some voice may sound higher in pitch, or simply "brighter", in presence of higher frequency resonances.

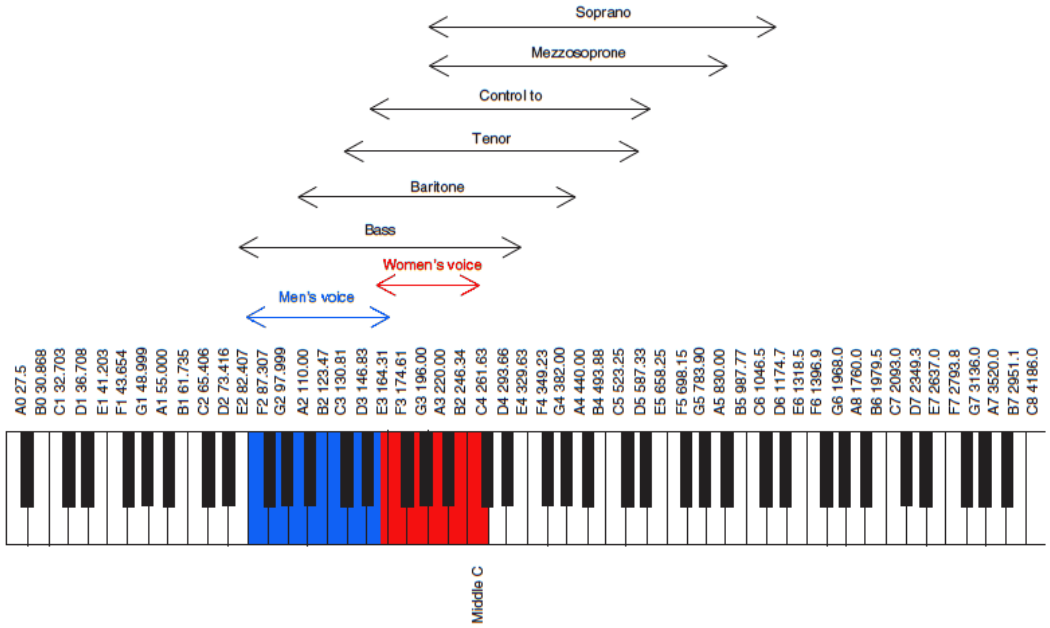

FIGURE 1.25: Some typical F0 ranges in different kinds of voice.

The rate of vocal fold vibration, and therefore the pitch of the voice, depends on the mass and on the stiffness of the vocals folds. Especially, the relationship between the mechanical parameters and the vibration rate is proportional to the square root of the ratio of the stiffness to the mass [42]. Moreover, independently of its mass, a vocal fold increases its vibratory frequency, and hence the pitch of the produced tone, when it is stretched and becomes stiffer. Equivalently, a lower vibration frequency characterizes more massive vocal folds instead of shorter and/or thinner folds. This is the reason why men normally produce lower-pitched voices than women or children [60]. Any speaker can control the rate of vocal fold vibration by performing a stretching of the vocal folds, in this case the rate will increase, or relaxing them, and hence decreasing of the rate.

In the literature there is still not a complete agreement on the manner in which F0 is physiologically controlled. Within a single individual the weight of vocal folds does not seem to change, except if the speaker is suffering from some disease or excess mucous. However, the contracting action of the thyroarytenoid and cricothyroid muscles modifies the $effective\ mass$ of the vocal folds. For instance, at a constant level of cricothyroid contraction, the thyroarytenoid contraction induces a shortening and compacting of the vocal folds resulting in an effective mass increase and in a F0 decrease. Therefore, it

is not possible to address changes in mass as the only factor influencing F0. Tension changes seem to contribute similarly to the modification of the vibratory rates as changes in vocal folds thickness [39, 42, 61]. The stiffness of the vocal folds can be changed by the cricothyroid muscle, that stretches them and that can act independently of the thyroarytenoid muscle. When the cricothyroid contracts, but the thyroarytenoid does not, the vocal folds are lengthened, their effective mass is decreased, the stiffness is increased and F0 increases.

The scenario is complicated by the ability of laryngeal muscles to contract independently and by the complex layered structure and the elasticity of the vocal folds. For instance, the isolated contraction of the thyroarytenoid muscle decreases the length of the vocal folds, and increases the stiffness of the body of the vocal folds, but decreases the stiffness of the cover due to a shortening of the body. The result is a small decrease in F0, but a large increase in loudness [62]. A joint contraction of both cricothyroid and thyroarynoid muscles does not produce any F0 change, since the two actions oppose each other. Finally, an increase of subglottal pressure produces an increase in F0. In fact, increasing subglottal pressure without an attempt to adjust the tension of the vocal folds increases the amplitude of the vocal fold excursion, stretching them by virtue of their greater lateral motion, and hence increases frequency. Such an effect is more appreciable at low frequencies, when the vocal folds are moderately slack, and can partially explain why vocal pitch tends to increase during yelling. Controlling all these aspects separately is possible, but difficult. In fact, normally it requires vocal training.

1.5.2 Intensity and loudness

The measurement of intensity, amplitude and loudness are different. Sound power per unit area is the definition of *intensity*, that is normally measured by means of an assessment of the sound intensity in the air as units of watts per square millimetre or centimetre. *Amplitude* is a measure related to the displacement of air molecules from rest when the sound waveform is visualized on a display. Finally, *loudness* is a psychoacoustic description of the relationship between a sound's intensity and the magnitude of the resulting auditory sensation. Amplitude can be thought as the perceptual correlate of the acoustic signal intensity [7]. Similarly to the relationship between pitch and F0, the relationship between loudness and intensity is not a linear function, but approximately logarithmic. Commonly, sound intensity is measured relative to a standard threshold of hearing intensity on decibel scale. Such scale ranges from 0, threshold of hearing, to 140, jet aircraft noise at a distance of 120 feet) [63].

The association between subglottal pressure and acoustic intensity has been investigated. Changes in intensity are often attributed primarily to modifications in subglottal pressure [64, 65]. As described in aerodynamics, the relation among pressure and resistance at the glottis and airflow rate is described by (equation 1.1):

$$Pressure = flowrate \times resistance \tag{1.1}$$

Consequently, loudness can be increased via pressure increasing in laryngeal resistance or flow rate, or both. The flow rate is changed by a adjusting respiratory effort. More pressure from the lungs should induce more airflow if glottal adjustments are kept constant. A change in resistance generates necessarily a modification in the laryngeal adjustment.

The interaction of these variables is complex. Especially, though intensity seems to increase with laryngeal resistance [66–68], some observations demonstrate that there is no persistent association between increasing intensity and laryngeal muscle activation, that may increase resistance [69, 70]. Experimental studies have largely been limited to excised larynxes, which can not be considered a complete model, since cannot take into account the contributions of the active contraction of the thyroarytenoid muscle. Generally, experimental data obtained by manipulating the thyroarytenoid muscle activity directly are lacking. Some results seem to indicate that large differences may be found between speakers in the precise balance between respiratory and laryngeal factors in regulating intensity [67]. Experimental studies aiming at the simultaneous modification in vivo of muscle stimulation, airflow, subglottal pressure and laryngeal resistance, and hence controlling for F0 changes, are necessary to observe these complex interactions among variables. Though speakers are able to control vocal loudness without much thought or effort, much more research is required to fully understand the way in which they achieve this.

1.5.3 Quality and phonation types

Another perceptual vocal characteristic is voice *quality* or *timbre*. It is not easy to define quality, since it does not have a fixed acoustic correlate. Anyway, if voice production is considered, quality in its narrow sense can be related to changes in the tension and mass of the vocal folds, to the symmetry of vibration, to the strength with which they are held together (medial compression), and to the measure of subglottal pressure [39]. In fact, it is possible to effectively perceive changes in the sound of a voice due to specific modification in the vibrational way of the vocal folds. In the following, some of the so-called *phonation types* will be described.

Normally, people are characterized by a kind of phonation that is known as *modal phonation*. In this definition, modal is used in the statistical sense of "modal". In addition, people can also produce a range of *nonmodal* phonation types. Among these phonation types, the way vocal folds vibrate can vary largely. For instance, $falsetto$ (Figure 1.26 a) is generated by a different vibratory mode of the vocal folds and it occupies the upper frequency limits of a speaker's vocal range. In fact, in this phonation type, vocal folds are put in vibration in a way which only the free borders can come into contact, while the rest of the folds relatively fixed. The vocal folds during falsetto appear long, stiff, very narrow, and may be somewhat bow- shaped [7]. In $period-doubled$ or *subharmonic phonation*, cycles alternate in a repeating long- short- long- short or large- small- large- small pattern (Figure 1.26 b) [7]. Instead, in $vocal\ fry$, that can be associated with perceived *creaky voice* [71], the vocal folds open and close abruptly, but the closed part of the cycle is longer (Figure 1.26 c) [7]. Both period doubling and vocal fry can occur very commonly in normal speech. Both have a unique timbre that listeners can easily identify. As opposed to, from period doubling and vocal fry, that they are generated by qualitative changes in vocal fold vibration, *breathy* nonmodal phonation (Figure 1.26 d) seems to form a continuum from modal phonation (not breathy at all) at one extreme, through *whispery* or breathy or *murmured* phonation (somewhat breathy), to whisper at the opposite extreme [7]. Passing from modal phonation to whisper phonation, across this continuum, vocal folds are closed more and more gradually. This enables generating less high-frequency acoustic energy. Moreover, at the end of each cycle the vocal folds may not be completely closed. Therefore, the voice could also be mixed with a unmodulated airflow through the glottis, resulting in noisy voice. During whisper, vocal folds are able to vibrate only slightly or not at all, and the turbulence, that emerges as air rushes through the partially-closed glottis, generate alone the acoustic energy. This is the reason why peoples need a big breath to whisper loudly or for a prolonged time. Period doubling, vocal fry, and breathy voice are very common in modal and daily phonation, since they can may have communicative functions [7, 71].

Sometimes, changes in vocal quality, in a broader sense, can also be reflected in changes in the resonant frequencies of the vocal tract. Changes in vowel quality are associated with shifts in the frequencies of the lowest three or four formants. On the contrary, the frequencies of the higher formants are associated by some authors with the "personal quality". Formants can be very informative about speakers. In fact, listeners are usually quite sensitive to their changes, and they can reliably perceive modification of as little as $4-8$ % in formant frequencies [58, 72].

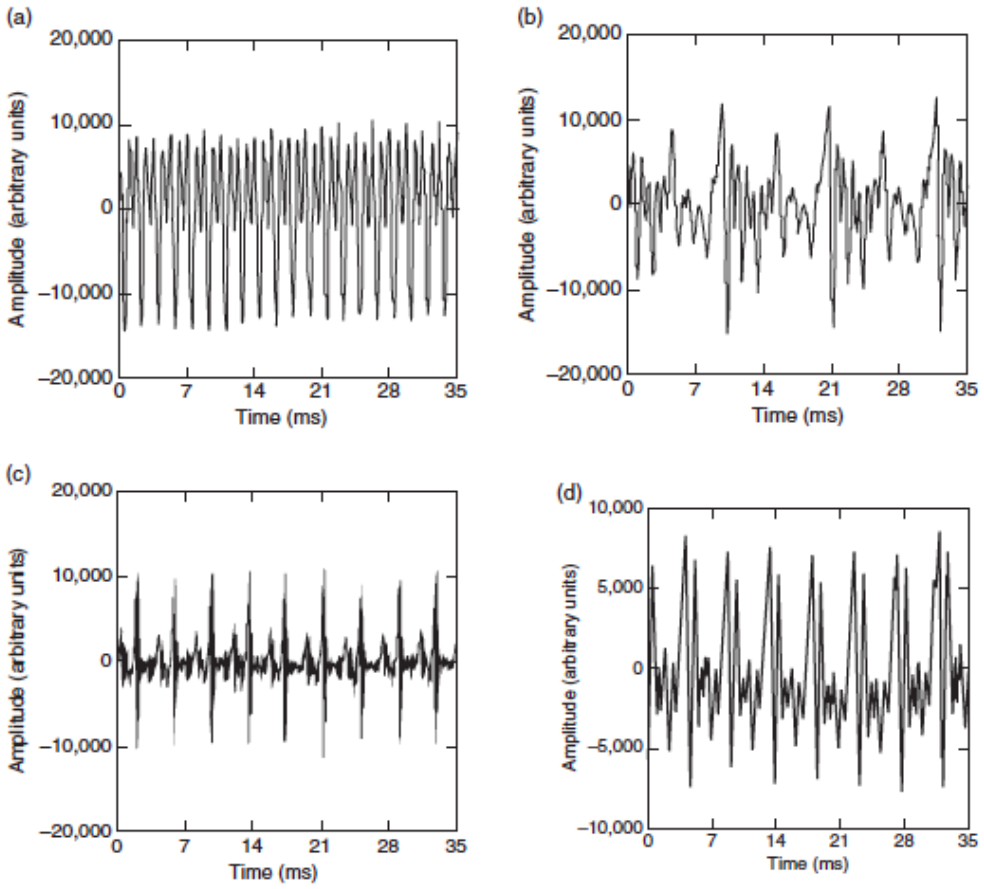

FIGURE 1.26: Time domain waveforms of some nonmodal phonation types. In (a) waveform of falsetto, in (b) the waveform of period-doubled phonation, in (c) the waveform of vocal fry, and in (d) the waveform of breathy voice are reported [7].

1.5.4 Individual voice quality

The understanding of the physiology beneath voice production is a good starting point to understand the way in which speakers can have a personalized, individual voice quality, and how people can vary their voice. The orchestration of the respiratory, laryngeal and vocal tract movements characterize the phonatory behavior. More than 100 different muscles are coordinated with regard to this aim. If people want to vocalize some sound, they have to control an appropriate amount of breath, create a air stream through the glottis (which must be properly configured and controlled in real-time), and then continuously they have to move simultaneously the jaw, tongue, velum and lips to adjust the airstream. A modification of the respiratory driving power can alter loudness. Changes in some laryngeal parameters can result in a modification of the mean frequency of speech, of the range of frequencies used, of the shape of the F0 contour, of the shape and/or timing of glottal pulses, of the phonation type, loudness, and so on [7]. Modification in the acoustic resonances can be generated by changes in the shape of the vocal tract, resulting in a production of different speech sounds or different accents. At

the end, all these factors can be altered dynamically over time, into a variety of vocalizations including talking, yelling, singing, sighing, laughing, humming, cursing, and reciting memorized material, among others [7]. Peoples are able to change the amount of variability in a voice, producing different sounds in succession.

Notwithstanding speakers are able to modify their voices over wide ranges of loudness, pitch and quality, their anatomy and physiology limit the effective range of sounds that can be produced [73, 74]. For instance, the length and the mass of the vocal folds limit the range of vocal folds vibratory rates. Speakers can span F0 over several octaves, but there are absolutely some individual limits that they cannot reach. Furthermore, speakers can modify the formants they use in a great variability of ways, but such variation are limited and this limit is determined by an individual's underlying vocal anatomy.

According to Laver [75], voice quality can be thought as the result of two main sources. On the one hand the anatomy and physiology of the speaker determine the width of the potential range of operation. On the other hand, the long-term muscular adjustments, or *settings*, of the larynx or the supra-glottal vocal tract restrict such potential range. In his model, Laver stated that some typical patterns identified from speech records may be explained as a deviation from a specific vocal tract configuration, or *neutral setting*. In this setting, articulatory organs show equilibrated muscular tension throughout the vocal tract during phonation. The identification of such settings was based on the Long Term Average Spectrum (LTAS). In fact, in line with this model, a setting is an average state of the vocal tract. Thus a given voice quality can be imagined as the acoustic result of a specific average articulatory configuration.

A primary distinction, among the sources of differences in vocal quality, is made between *organically-* based differences between speakers, and differences that are due to *learned* or *habitual* behavior [7, 76, 77]. Hence, organically- based differences depend on the relatively unchangeable, physiologically-based characteristics. Mean F0 and formants are also included. From a theoretic point of view, since they are related to the physiology, in adult speakers such parameters could be very stable over time, offering therefore some good indices of speaker's identity, sex, and age. On the contrary, learned differences between speakers depend on the experience and they include accent, speaking rate, intonation contours, habitual F0, specifics of voice quality, and so on. Though they are partially under the speaker's control, such characteristics are also often stable, since people tend to send signal group membership and important personal attributes.

These two kind of characteristics, organic and learned, are the reason of why family members sound similar. Heredity causes family members to have similar laryngeal and vocal tract shapes and sizes. Family is also the environment in which speech is first

learned. Siblings often grow up in the same linguistic community, and tend to acquire the same vocal patterns. Physiological, dialectal and idiolectal, i.e. unique speech pattern belonging to a single person, similarities combine themselves to create a family voice that allows in a relatively easy way to detect its members from the others, but on the other hand make it harder to distinguish the different members, especially if they are close in age [7].

Until now, the sources of differences between speakers, $inter-speakers\ variations$, have been discussed, but the difference within a single speaker in speaking style, $intra-speaker\ variability$ is also important. Every speaker sounds different from day to day, and from time to time within a day. For instance, many speakers are hoarse in the morning, and most speakers vary their voices across different emotional state or with fatigue. Many authors asserted that such $intra-speaker$ variability is small with respect to $inter-speaker$ variability. Such assumption can simplify considerably the work when discussing speaker recognition, because it reduces the task [7]. Anyway some studies suggested that differences within the same speaker may sometimes be as large as those between speakers [78, 79].

1.6 Speech in the time and frequency domains

The speech signal is a slowly time varying signal. Hence, this means that if examined over a sufficiently short period of time, $5 \div 100ms$, its characteristics can be considered fairly stationary [14]. Anyway, the speech signal characteristics change as consequence of the different speech sounds being spoken over periods of time of the order of 1/5 seconds or more.

Speech events can be classified or labelled in many different ways. The simplest takes into account three different states. The first is *silence*, during which speech is not produced. The second state is *unvoiced* in which the vocal folds are not in vibration, resulting in a speech waveform is aperiodic or random. Finally, the third state is *voiced*. In the voiced state, the vocal folds are tensed and vibrate periodically when air flows from the lungs. Its speech waveform is pseudo-periodic.

The distinction between well-defined silent, voiced and unvoiced frames is not exact. Often, it can be difficult to distinguish a weak unvoiced sound, for example [f] or [h] from silence, or a weak and voiced sound, like a [v] or [m] from an unvoiced sound or even silence. Normally it is possible to locate and segment speech signals into these three different categories with a precision of several milliseconds [14]. Errors in boundary locations are thus negligible for many applications.

Another way of characterizing speech sounds is by a spectral representation. In a spectrogram, the sounds are described, over time, via relative intensity in different frequency bands. An example is in Figure 1.27 [6]. The first panel of Figure 1.27 reports the time waveform of the speech signal, the second a *wideband spectrogram*, and the third the *narrowband spectrogram*. The wideband spectrogram was estimated with a window length equal to 10 ms. Such a window length is of the order of the length of a glottal cycle in voiced intervals. In voiced intervals, the spectrogram reports vertically oriented striations. The reason of these is that the sliding window includes alternately mostly large amplitude samples and then mostly small amplitude samples. Hence, each individual cycle can be located in time, but the frequency resolution is low. This is the reason why this kind of spectrogram is called wideband. Differently, the narrowband spectrogram was computed with a window length equal to 40 ms. Such a window length comprises several cycles of the waveform during voiced intervals. Therefore, the spectrogram can no longer displays vertically oriented striations. This kind of spectrogram also is not sensitive to quick time variations, but the frequency resolution is much better, resulting in striations that tend to be horizontally oriented. Unvoiced sounds are primarily notable by means of their high-frequency energy, while silence is essentially characterized by the lack of any spectral activity.

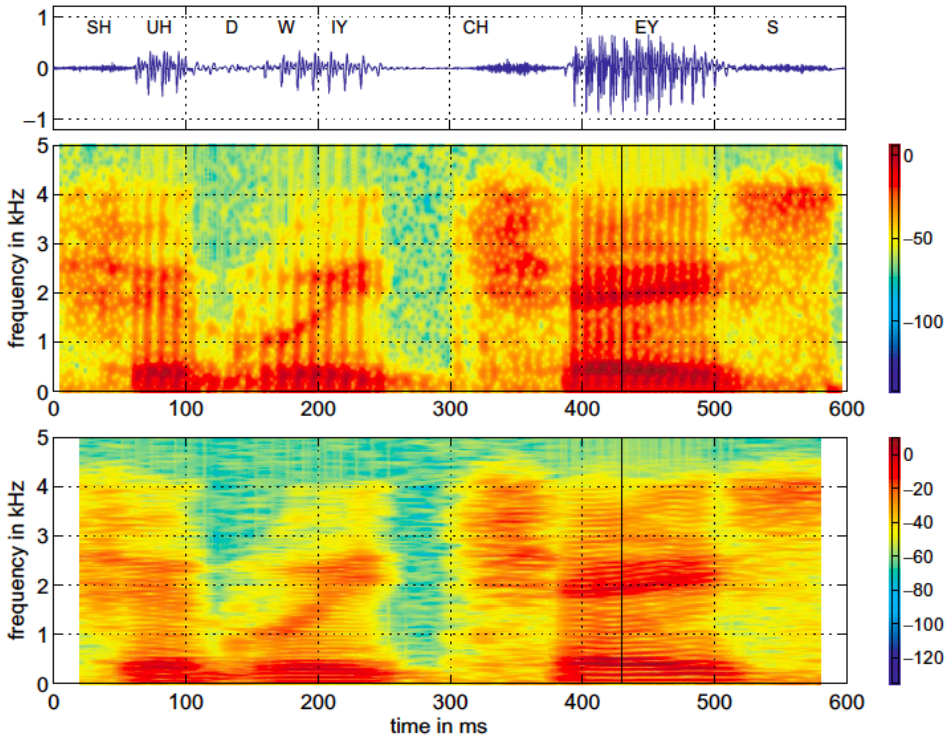

FIGURE 1.27: Spectrogram for a speech signal. The spoken phrase is "Should we chase" [6].

Another way of labelling the time-varying speech signal characteristics is by means of a

parametrization of the spectral activity based on a model of speech production. Since the human vocal tract can be mainly modelled as a tube, or a set of tubes, of varying cross-sectional area where the exciting source is located at one end, i.e. the pulsatile airflow, or at a point along it, i.e. the noise, the acoustic theory asserts that such a system can be described in terms of resonance frequencies or formants. Spectrograms can be reliably used to estimate formant frequencies.

1.7 The phonetic alphabet

To transmit information, and thus to be a reliable medium of communication, language has to be composed by a finite and mutually exclusive number of distinguishable sounds. This means that language has to be constructed of basic and fundamental linguistic units. A modification, within a utterance, of one of these may result in a change of meaning. The acoustic manifestation of these basic units can have a great variability, but every modification designates the same linguistic element in a listener who is skilled in that language. Such a basic linguistic element is called *phoneme* [80] and its multiple acoustic realisations are called *allophones*. Thus, the phonemes may be considered as a kind of code uniquely related to the articulatory gestures of the considered language. Again, the allophones may be thought as expression of the acoustic freedom permitted in the coding of a specific phoneme. It is interesting to notice that such a freedom does not only depend on the related phoneme, but also on its position in the utterance. For every language, and also for every dialect of the communicators, the set of code symbols used in speech, and their statistical properties, are defined.

The first step, performed by a linguist who is studying an unknown language, is always the phonetic transcription. In this transcription, every perceptually-distinct sound is reported with a different code symbol. Then, the linguist's subsequent step is the effort to relate the code to the behavior, in order to understand if some acoustically distinguishable sounds belong to the same phoneme. This means that all the sounds that do not belong to a different phoneme are grouped in the same class. Therefore all these sounds are different, but their difference is not relevant to the meaning of the linguistic units they belong to, it is just a convention of the language. Different languages usually have their own speech characteristics. Some of these can be phonemically distinct in one language, but not in another one. For instance, a simple change of pitch of a vowel can be reflected in a modification of the meaning in many East Asian and Western African languages. This is not generally true in European and Middle Eastern languages. Moreover, in Bantu languages of southern Africa, such as Zulu, tongue clicks and lip smacks are genuine phonemes.

From these considerations, it results that speech is, in some sense, discrete, though a proper representation of the emitted sound pressure wave, and regarding the spoken speech, is continuous. Connected speech is produced by a continuous motion of the vocal apparatus from target to target. During this motion, a continuous adjustment of the vocal tract configuration is performed as well as of its modes of excitation. In this frame, a given phoneme is produced by a momentary and particular configuration of the vocal tract. The precision with which a phoneme needs to be articulated is greatly influenced by the statistical constraints of the language. Sometimes, a phoneme can be simply signalled by making a vocal gesture in the direction of the normal configuration. In fact, the relations between speech sounds and vocal motions are not unique, as demonstrated by the compensatory articulation of ventriloquists and the mimicry of parrots.

Notwithstanding the great variability of the vocal apparatus in connected speech, and its continuous nature, humans are able to subjectively segment speech into phonemes. Transcriptions of connected speech events can be written by phoneticians, who developed a proper phonetic alphabet for this aim. The international phonetic alphabet (IPA) provides code symbols for representing the speech sounds of most of the major languages of the world. Several different levels of precision can be observed in linguists' transcriptions. Two phonemes are considered different if only they are able to change the meaning of a word by switching them. A "phonemic" transcription is a transcription made in terms of phonemes. Usually it is enclosed in slashes / / [81]. On the other hand, the IPA provides also many unused acoustic distinctions, able to change the meaning of a word, in any given language. A transcription that specifies every allophonic or sub-phonemic distinctions is called "phonetic", and usually is enclosed in brackets [ ]. Normally, manner and place of production of the speech sounds contribute to their classification. For instance, the description of the position of the tongue hump along the vocal tract and the degree of the constriction are used to describe the vowel sounds.

1.7.1 The Vowels

Perhaps the most interesting category of sounds in English are the vowels. They are not so important in written text. In fact, if a text is deprived of every vowel, it is usually still readable, while if it is deprived of its consonants, the meaning usually is lost.

Vowels sounds are produced by pseudo-periodic pulses of air, caused by the vibration of the vocal folds, that excite the vocal tract. During this sound production the vocal tract is essentially set in a particular shape. The resonant frequencies of the tract, i.e. the formants, are controlled by the manner in which the cross-sectional area varies. The positions of the tongue and of the jaw, but also the velum, determine the vowel sound. More in detail, the length of the vocal tract affects the frequency locations of all vowel

formants accordingly to a simple inverse proportionality rule. In Table 1.1 some rules, relating typical vocal tract shapes and formants, are reported.

TABLE 1.1: Rules Relating Formant Frequencies and Vocal-Tract Characteristics for the Vowel Sounds [12].

Length Rule: The average frequencies of the vowel formants are inversely proportional to the length of the pharyngeal-oral tract (i.e., the longer the tract the lower its average formant frequencies).
F_1 **Rule-Oral Constriction:** The frequency of F_1 is lowered by any constriction in the front half of the oral section of the vocal tract. The greater the constriction the more F_1 is lowered.
F_1 **Rule-Pharyngeal Constriction:** The frequency of F_1 is raised by a constriction of the pharynx, and the greater the constriction the more F_1 is raised.
F_2 **Rule-Back Tongue Constriction:** The frequency of F_2 tends to be lowered by a back tongue constriction. The greater the constriction the more F_2 is lowered
F_2 **Rule-Front Tongue Constriction:** The frequency of F_2 is raised by a front tongue constriction. The greater the constriction the more F_2 is raised.
Lip-Rounding Rule: The frequencies of all formants are lowered by lip-rounding. The more the rounding the more the constriction and subsequently the more the formants are lowered

The duration of the vowel is usually long, when compared to the lengths of consonants. Normally, vowels are placed among the phones of largest amplitude. Vowels can vary widely in duration (typically from $40-400$ ms).

In the literature, many methods aiming at the characterization and classification of vowels have been reported. Some are based on the articulatory configurations that are required to produce that particular sound. Others, instead, investigate their typical waveform plots or their typical spectrogram plots. In Figures 1.28 and 1.29 the articulatory configuration for some typical vowel sounds [12] are displayed with their corresponding acoustic waveforms and their corresponding vocal-tract magnitude spectra. Usually, vowels are conventionally classified by means of the articulatory configurations in terms of tongue hump position. This, in fact, can be usually placed in the front, mid or back. Another configuration parameter is the tongue hump height, i.e. high, mid or low. In such a frame, tongue hump is considered to be the summit of the tongue at its narrowest constriction within the vocal tract. According to this criteria, the vowels [i], [I], [æ], [ɛ] and [a] are classified as front vowels, while the vowels [ɑ], /ʌ/, /ɔ/, /U/, /u/ and /o/ are back vowels. The front vowels are characterized by a high-frequency resonance. In the mid vowels, instead, a balance of energy over a broad frequency range can be detected, while in the back vowels a low-frequency F1 is characteristic. Certainly, these behavior can be also recognized in their spectrogram. In fact, a relatively high second and third formant frequencies can be detected in the front vowels. Mid vowels report well-separated and balanced formants, and finally, back vowels, especially [u] do not show any energy contribution beyond the low-frequency region. In this case first

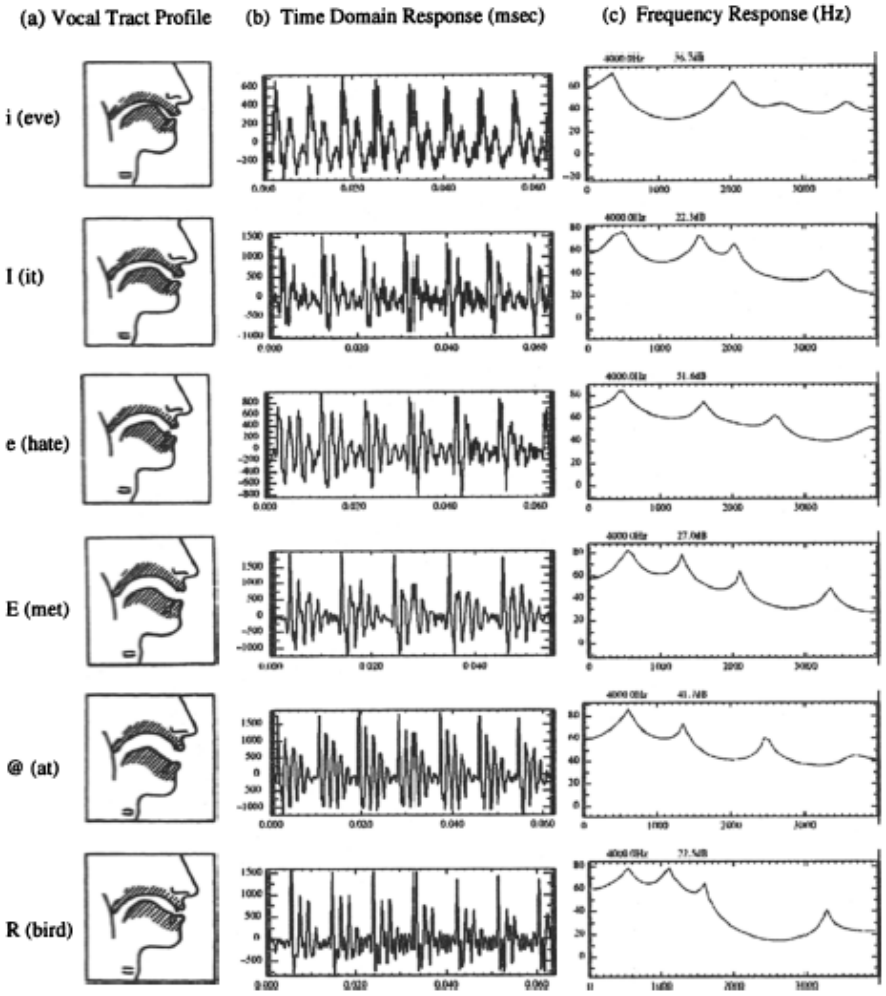

FIGURE 1.28: Vowels in American English. In the first column, (a), some schematic profiles of vocal-tract are reported, in column (b) some typical acoustic waveforms, and in column (c) the corresponding vocal-tract magnitude spectrum for each of the reported vowel are represented [12].

and second formants are very low. Average formant locations for vowels in American English can be observed in Figure 1.30.

The variability of vowel pronunciation among people, across gender, age, regional accents and other variable characteristics make the concept of "typical" vowel quite unreasonable. An example of such variability can be seen in the plot (Figure 1.31), made by Gordon Peterson and Harold Barney, of the first and seconds formant frequencies [13]. For a given vowel sound, a great difference can be detected in the two studied formants across speaker. In addition, it is important to highlight that there is a overlap between the formant frequencies for different vowel sounds for different speakers. The ellipses drawn in the Figure are representing the gross formant configuration area for each investigated sound. Noticeably, it is not sufficient to estimate formants or spectral peaks to reliably characterize and classify vowel sounds. Some speaker-dependent normalization

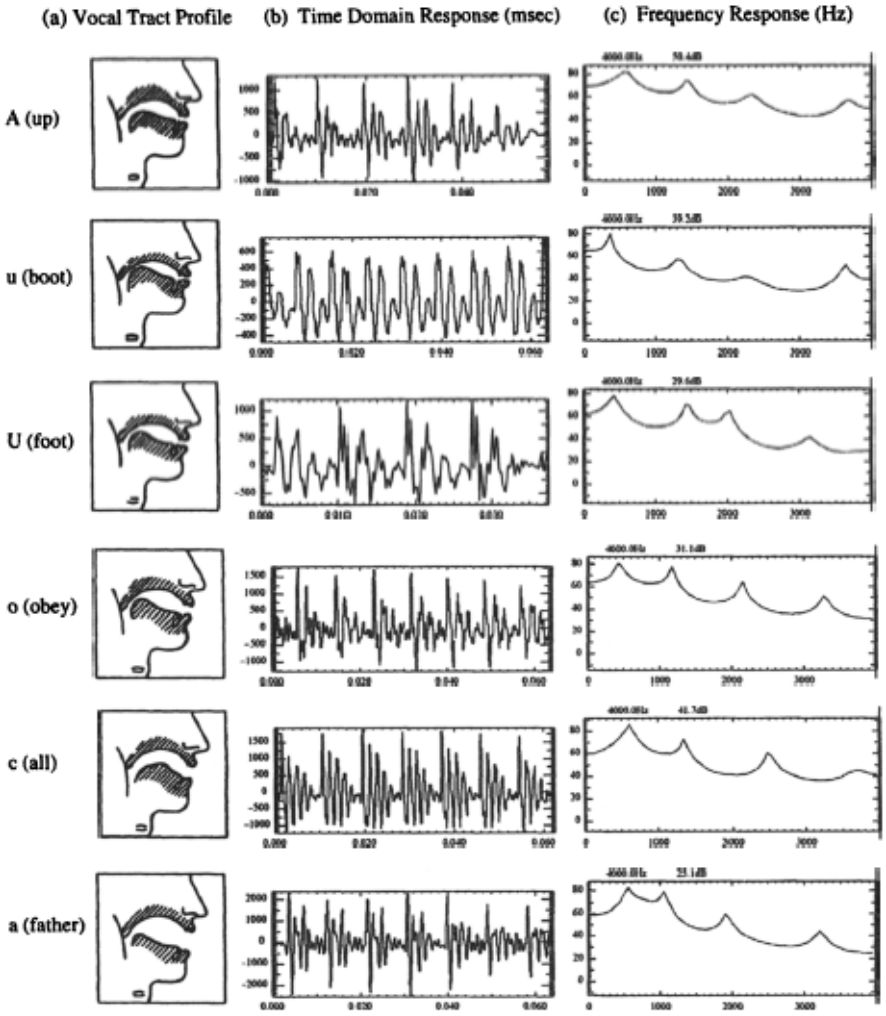

FIGURE 1.29: (continued) [12]

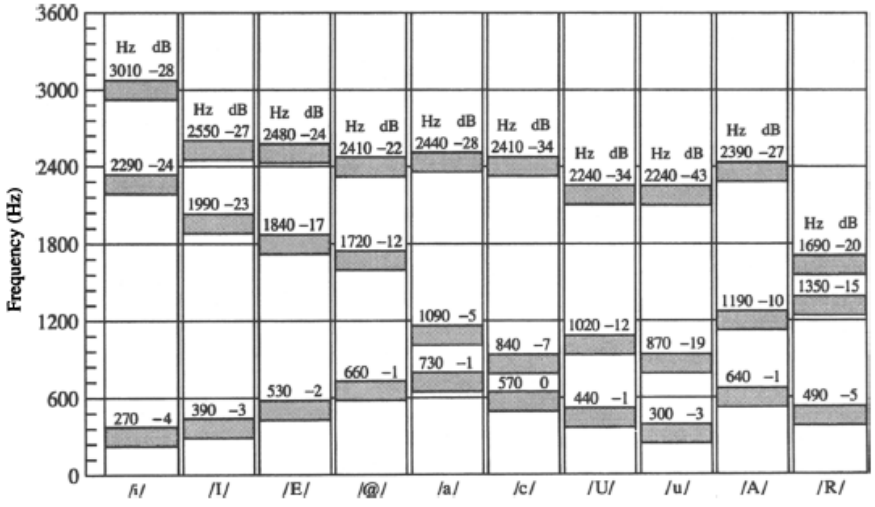

FIGURE 1.30: Average formant locations for vowels in American English [12, 13].

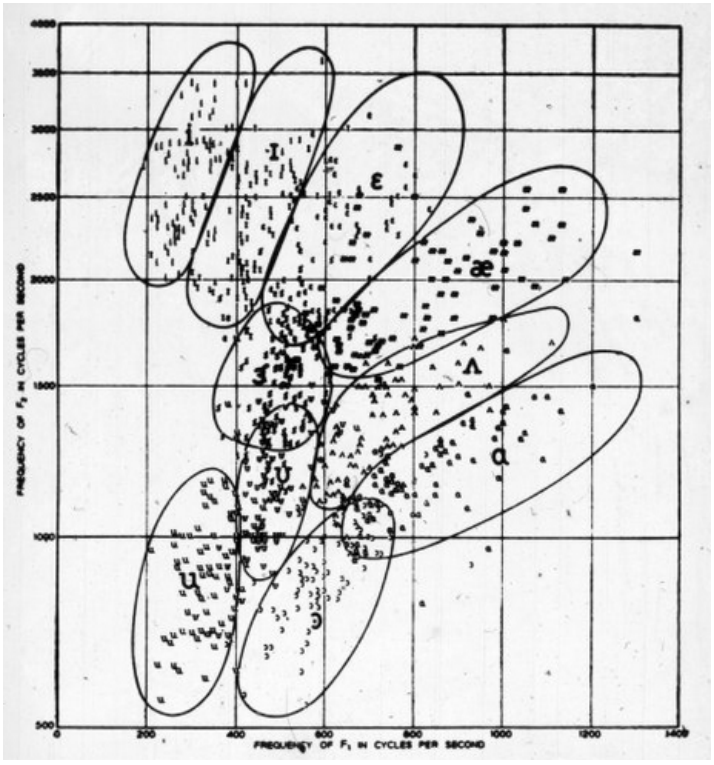

FIGURE 1.31: Frequency of second formant versus frequency of first formant for ten vowels by 76 speakers [13].

has to be performed to account for the variability in formants and the overlap between vowels.

A useful way of representing vowels is shown in Figure 1.32, i.e. the so-called vowel triangle. In this Figure each vowel is represented by a centroid in the formant space, under the hypothesis that this centroid represents at the best the average behavior, and does not represent the variability across speakers. The vowel triangle is formed at its extremes by the formant locations of the vowels [i] (low F_1 and high F_2), [u] (low F_1 and low F_2), and [a] (high F_1 and low F_2).

1.7.2 Diphthongs

Different definitions of diphxthongs may disagree. An acceptable definition is the one that defines a diphthong as a gliding monosyllabic speech sound that starts at or near the articulatory position for one vowel and moves to or toward the position for another [14]. Usually, the first target vowel is longer than the second, even if the transition between them is still longer [82]. In line with such a definition, it is possible to recognize six diphthongs in American English, namely [ai] (as in buy), [au] (as in down), [ei] (as in bait), [ɔi] (as in boy), [ɔu] (as in boat).

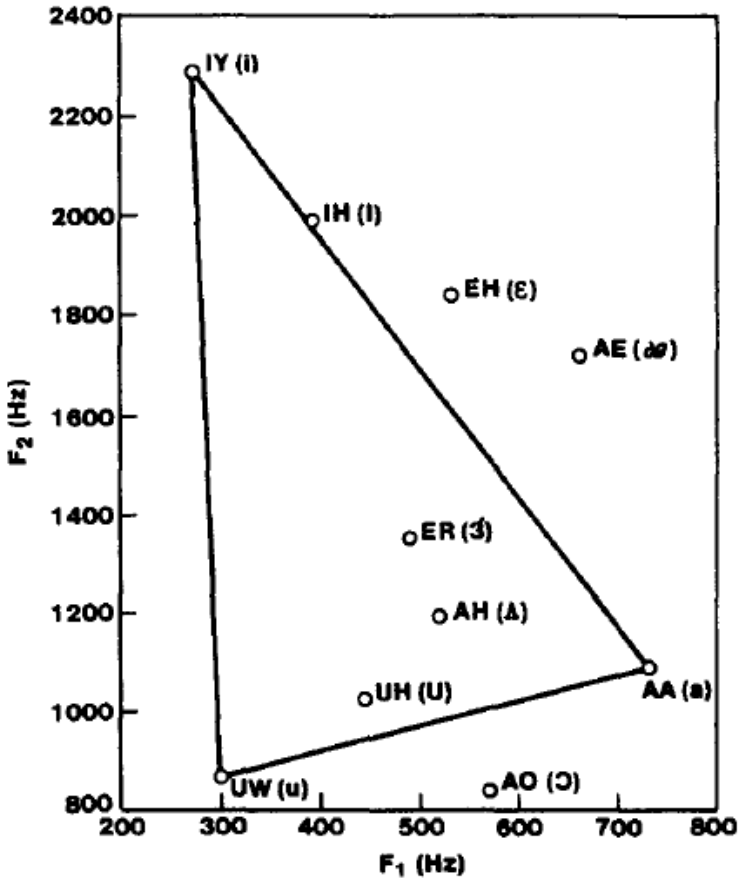

FIGURE 1.32: The vowel triangle with centroid positions of the common vowels [14].

To highlight the time-varying spectral characteristics of diphthongs one may study them in the F1- F2 plane. In Figure 1.33 the arrow indicate the direction of motion of the formants over time. The vowels average positions are reported by means of dashed circles. For these reasons, diphthongs can be characterized by a time-varying vocal tract area function that varies between two vowel configurations.

1.7.3 Semivowels

The semivowels, [w], [l], [r] and [ɥ], (Figure 1.34) are a group of sounds that is not easy to characterize. The group name derives from their vowel-like nature. Normally, they are characterized as a gliding transition between adjacent phones. This means that their acoustic behavior is strongly dependent on the context in which they are placed. Their transitional nature make them similar to diphthongs, but their role is consonantal.

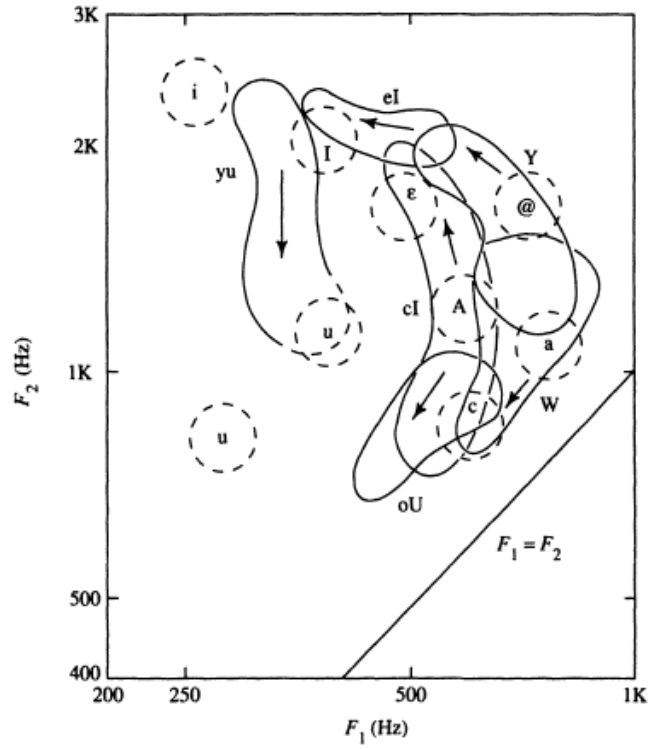

FIGURE 1.33: Movements of F1 and F2 for some diphthongs in American English [12, 15].

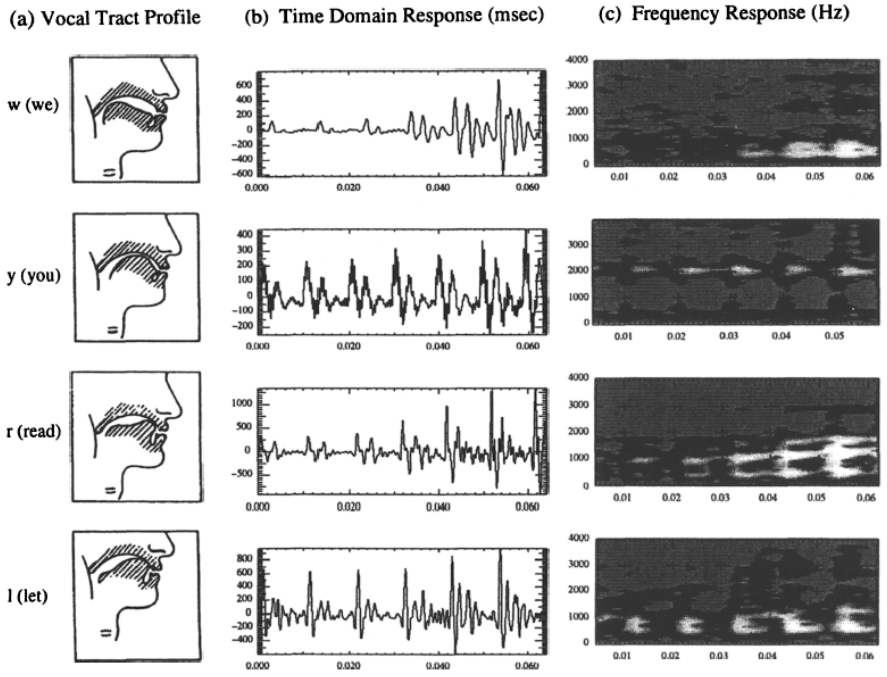

FIGURE 1.34: Semivowels in American English. In the first column, (a), some schematic profiles of vocal-tract are reported, in column (b) some typical acoustic waveforms, and in column (c) the corresponding vocal-tract magnitude spectrum for each of the reported semivowel are represented [12].

1.7.4 Nasal Consonants

When a glottal excitation occurs concurrently with a closure of the vocal tract, the produced sound is a nasal consonant: [m], [n] and [ŋ]. When the velum is lower the constriction can occur at any point along the oral passage. In this configuration, the velum is lowered and thus the air can flow through the nasal tract. This results in a sound that is radiated at the nostrils. Though the oral cavity is constricted toward the front, it is still acoustically coupled to the pharynx. Therefore the mouth can serve as a resonant cavity, which causes anti-formants. They appear as zeros in the transfer function. In addition, nasal consonants and *nasalized vowels*, that are vowels that proceed or follow nasal consonants, are characterized by resonances that are spectrally broader, or more highly damped, than those of vowels. The place along the oral tract at which the closure occurs can be used to distinguish the three nasal consonants. For example, concerning [ŋ] such a constriction is just forward of the velum itself.

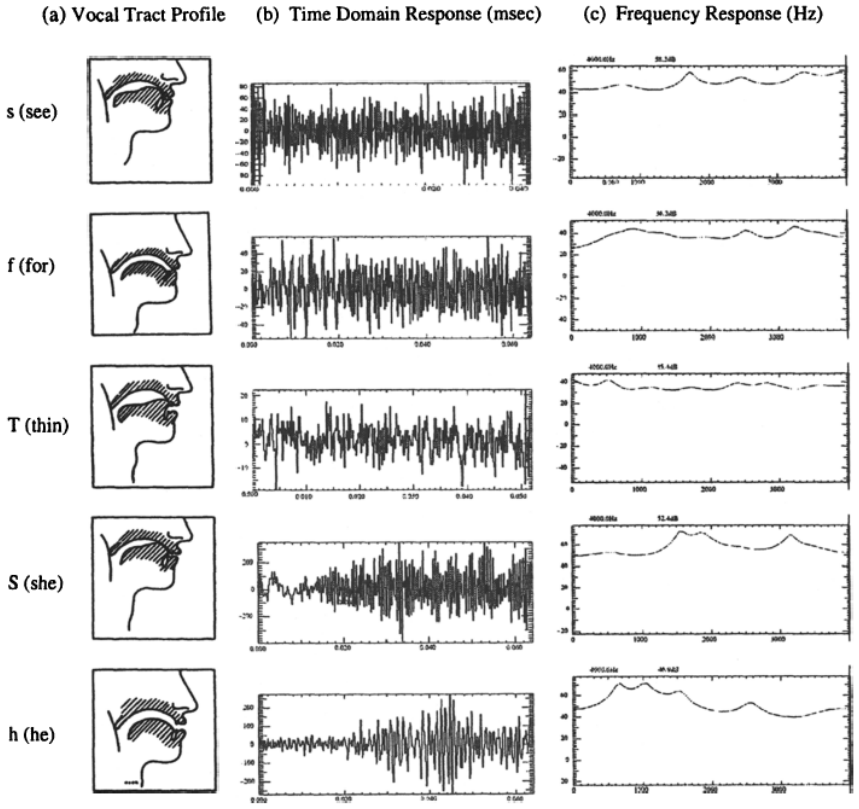

FIGURE 1.35: Unvoiced fricatives in American English. In the first column, (a), some schematic profiles of vocal-tract are reported, in column (b) some typical acoustic waveforms, and in column (c) the corresponding vocal-tract magnitude spectrum for each of the reported unvoiced fricatives are represented [12].

1.7.5 Unvoiced Fricatives

When the vocal tract is excited by a steady airflow, which becomes turbulent in a constricted region, unvoiced fricatives are produced. This constriction, located in the vocal tract or glottis, results in an unvoiced excitation. These sounds are, [f], [θ], [s] and [ʃ], are characterized by a different point of articulation. The fricative [f] is generated by a constriction placed near the lips. The [θ] has a constriction near the teeth. The constriction for [s] is placed near the alveolar ridge, while the one for [ʃ] is slightly more back than the one of [s]. The generation of this sound category involves a source of noise at a constriction, which splits the vocal tract into two cavities. The sound is radiated from the front cavity at the lips. The role of the back cavity, as previously seen for the nasal, traps energy and introduces anti-formants into the vocal output. From the waveform plots of the unvoiced fricatives, it is easy to notice the non-periodic nature of these sounds (Figure 1.35). The major constriction and its effect on low-frequency energy content is evident. In Figure 1.35 vocal-tract profiles, time waveforms, and vocal-tract frequency responses for unvoiced fricatives are illustrated.

1.7.6 Voiced Fricatives

If the unvoiced fricatives [f], [θ], [s] and [ʃ] are generated by an unvoiced excitation, then the fricatives, [v], [ð] , [z] and [ʒ], are their voiced counterparts since the place of constriction for each of the corresponding phones is mainly the same. But, they are different from their unvoiced counterparts since two excitation sources are involved in their production. One excitation is the turbulence produced by the airflow in the neighbourhood of the constriction. The other is the pulsatile airflow at the glottis. In Figure 1.36 vocal-tract profiles, time waveforms, and vocal-tract frequency responses for voiced fricatives are illustrated.

1.7.7 Voiced and Unvoiced Stops

When an increasing pressure behind a complete closure somewhere in the vocal tract is suddenly released, the sound produced is a transient and non-continuant sound, known as a stop consonant. Voiced stops are: [b], [d], and [g]. To produce the [b] sound the closure occurs at the lips. The production of [d] requires a closure at the back of the teeth, while [g] requires one near the velum. When the tract is completely closed, no sound is radiated from the lips. But, a small amount of low-frequency energy, is radiated through the walls of the tract. It is called *voice bar*. This happens when the vocal folds are allowed to vibrate even if the vocal tract is closed at some point. The properties of

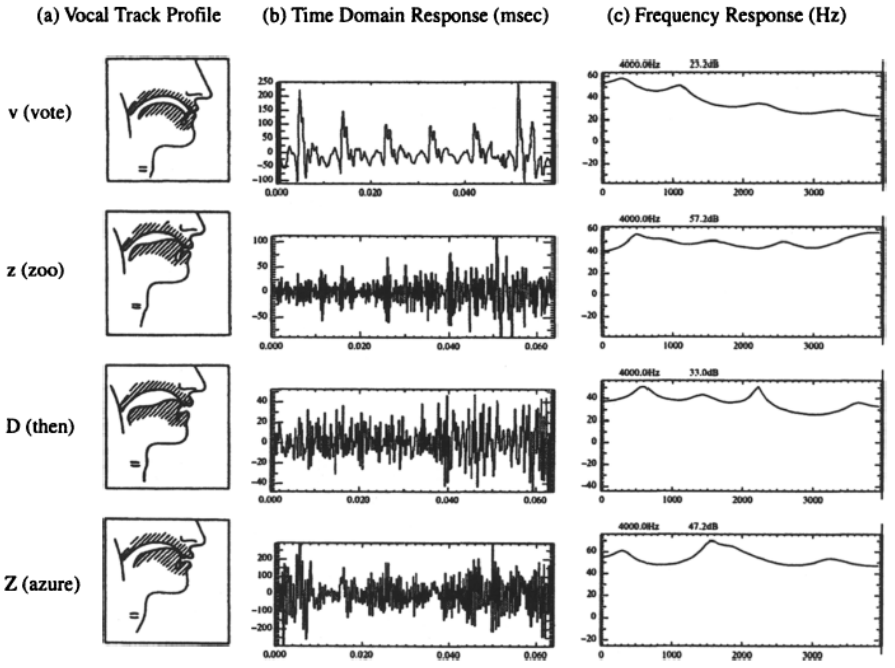

FIGURE 1.36: Voiced fricatives in American English. In the first column, (a), some schematic profiles of vocal-tract are reported, in column (b) some typical acoustic waveforms, and in column (c) the corresponding vocal-tract magnitude spectrum for each of the reported voiced fricatives are represented [12].

such stops sounds are highly influenced by the vowel that follows the stop consonant. The reason is that stop sounds are dynamical in nature. In English, most stops are not released if they occur at the end of a syllable. This phenomenon is produced by a reduced lung pressure that decreases oral pressure behind the occlusion. Therefore, the waveforms are not very informative regarding the place of articulation of a stop consonant. Few distinguishing features are shown by the waveform of [b], except for the voiced excitation and lack of high-frequency energy.

The unvoiced stop consonants [p], [t], and [k] are similar to their voiced counterparts [b], [d], and [g], except for when the pressure increases during the interval of closure of the tract, the vocal folds do not vibrate. Then, after the interval of closure, when the air pressure is released, a short fricative interval occurs, produced by a sudden turbulence of the exiting air, followed by an interval of aspiration, i.e. a steady airflow from the glottis that excites the resonances of the vocal tract, before voiced excitation begins.

In Figure 1.37 waveforms plots and spectrograms of the voiced stop [b] and the unvoiced stop consonants [p] and [t] are shown. It is important to notice that the "stop gap", or time interval during which the pressure is increased, is clearly detectable. In addition, the duration and frequency content of the fricative noise and aspiration noise vary greatly with the stop consonant.

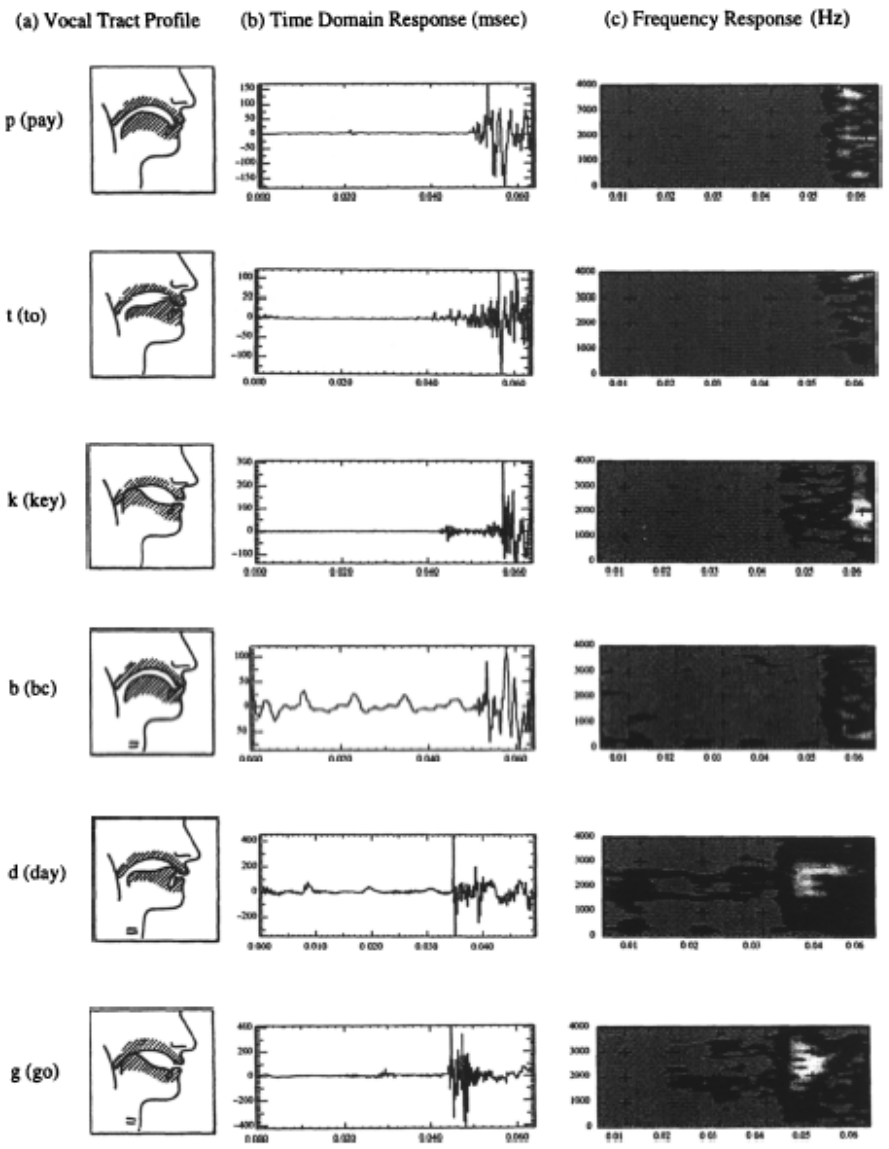

FIGURE 1.37: Voiced and unvoiced stops in American English. In the first column, (a), some schematic profiles of vocal-tract are reported, in column (b) some typical acoustic waveforms, and in column (c) the corresponding vocal-tract magnitude spectrum for each of the reported voiced and unvoiced stops are represented [12].

1.8 Prosody and Coarticulation

1.8.1 Prosody

The description given until now is focussed on the characterization of speech in terms of articulatory phonetics, i.e. the manner or place of articulation, and acoustic-phonetics, such as frequency spectrum and time waveform characteristics. During speech production, a complicated and rich sequence of articulatory movements are involved. Moreover, they are timed so that the shape of vocal tract is set in a particular way in the desired

phoneme sequence order [12]. Taking into account acoustic-phonetic arguments, the expressive use of speech depends on patterns of pitch, syllable stresses, and timing to form rhythmic speech patterns. Conventionally, *prosodic features* designate those aspects of speech related to tone and rhythm. Moreover, prosodic features are said to be *suprasegmental* since they often extend over more than one phonetic segment. The production of prosodic features involves both source factors and vocal-tract shaping factors. Subtle changes in the speech breathing muscles and vocal folds normally constitute the source factors, while movements of the upper articulators compose the vocal-tract shaping factors. The acoustic patterns of prosodic features are heard as systematic changes in duration, intensity, fundamental frequency, and spectral patterns of the individual phonemes.

According to many researcher, the time-varying vocal-tract transfer function is able to provide most of the information related to which phonemes are produced. Glottal source characteristics convey prosodic cues as *intonation* and *stress*, two of the most important prosodic features. Usually stress aims at distinguishing similar phonemes or highlighting a syllable or word. From a structuralistic point of view, four degrees of stress can be detected and normally distinguished: primary, secondary, tertiary, and weak.

The distinctive use of patterns of pitch or melody is known as intonation. Intonation is usually analysed by taking into account pitch patterns in terms of contours. For this aim, pitch range, height and direction of change are normally investigated. An important function of intonation is to signal grammatical structure. Noticeably, with a view to the development of natural sounding text-to-speech systems, intonation has to be seriously investigated and reproduced. In this frame, it is possible to assert that intonation plays a role similar to punctuation in writing. But, intonation has a much wider scope, including marking sentence, clause, or other boundaries, as well as contrasting grammatical sentence structure such as questions or statements. Another important role played by intonation is conveying secondary characteristics such attitude or emotion. In fact, emotions are marked by contrasts in pitch, even when other prosodic and paralinguistic features are also involved. Two parameters are involved in the modification of the glottal source in prosody. The first is the subglottal pressure, while the second is the tension of the vocal folds. These parameters can alter fundamental frequency, the source spectrum and the source amplitude. An increase in the lung pressure results in an increase in subglottal air pressure. This can increase the rate at which airflow pulses are produced at the glottis, resulting at the end in an increase of the fundamental frequency. Hence, subglottal air pressure augmentation can increase both pitch and loudness. Subglottal air pressure and pitch are approximately related by a straight line on a logarithmic scale of fundamental frequency versus subglottal pressure [83].

Subglottal pressure and intonation convey different stress patterns depending on what syllable the stress is placed and which word is pronounced in a statement or question. In Figure 1.38, one sees the patterns of subglottal air pressure and fundamental frequency for four sentences with the word "digest" [12, 83]. In Figure 1.38 (a) and (b) the word is spoken as a noun with stress on the first syllable, while the word is spoken as a verb with stress on the second syllable in Figure 1.38 (c) and (d). Noticeably, the stressed syllable is always related to higher subglottal pressure. Moreover, statements may present higher subglottal pressure than questions. With regard to the fundamental frequency, it shows a downward contour, from the beginning to the end of sentence, in statements. Stressed vowels are the longest. In statements, the position of the stressed syllable depends on whether the stressed syllable is a noun (high pitch in the stressed syllable) or a verb (higher pitch on the second syllable). If the same sentence is spoken as a question, normally the pitch contour rises from the beginning to the end of the sentence (Figure 1.38 (b) and (d)). Concerning noun in questions, the stressed syllable have neither increased duration nor pitch from the following syllable [12]. The reason is that the rising intonation contour of a question supersedes lexical stress requirements at the word level. Increased duration and higher pitch mark the stressed syllable in the final verb form. It is important to notice that, as demonstrated by this example, amplitude and duration of pitch contours corresponding to stressed and unstressed syllables depend on whether the intonation pattern is related to a question or a statement.

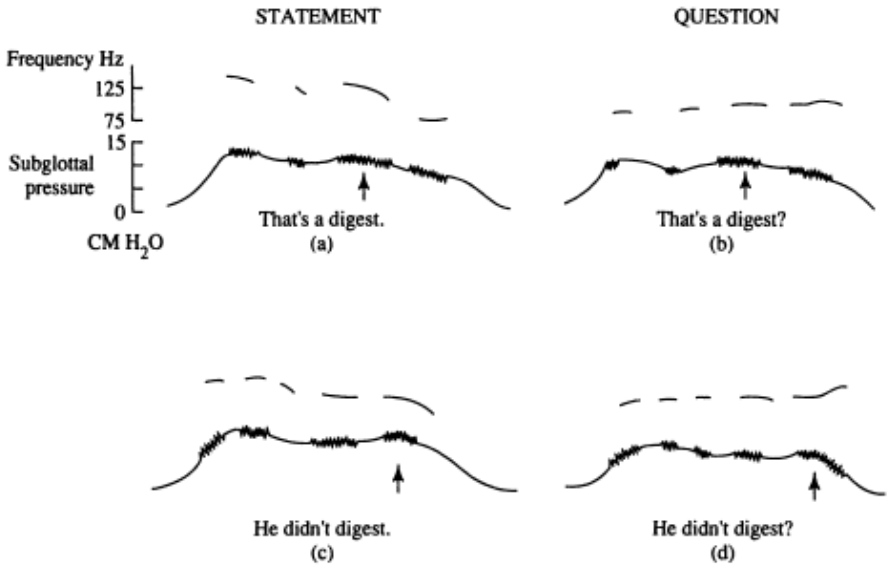

FIGURE 1.38: Relations between fundamental frequency contours and subglottal air pressure. Two statements and two questions with two different word stress patterns [12] are taken into account.

On the one hand, the discussed glottal source parameters contribute to stress and intonation. On the other hand, duration and its variation can also be used to convey prosodic

features. In fact, a syllable in the final portion of a breath group is usually longer that the other ones. Shorter vowels are normally part of unstressed words. Moreover, the duration of a spoken vowel can depend also on the consonant that follows. The duration of the vowel is affected by syllable stress.

1.8.2 Coarticulation

Moving the articulators in the vocal tract and the glottal source is required to produce speech. Phoneme articulations typically are overlapping in time, resulting in transitional sound patterns. Speech is not normally produced by means of quick rigid articulatory movement between uniform islands of stationary phone productions, but is produced by means of a smooth movement and to shape the vocal tract in accordance with the planned phoneme sequence [12]. Usually, the change in phone articulation and acoustics caused by the presence of other sounds in the same utterance is called *coarticulation*.

Some articulators can move fairly independently of one other. The degree and ease of movement depend on the muscles groups associated with each articulator, on their mass and on their position. Tongue and lips movements can overlap. In a vowel-consonant-vowel sequence, the tongue articulation depends on the target positions for each vowel and consonant. Speakers have considerable freedom of movement when each phoneme involves different sets of articulators.

An articulator may be displaced toward a position more appropriate for the following phone, in the absence of a strong conflict. This kind of anticipatory coarticulation is known as $right{-}left$ because the target influences the production of the phone. Moreover, the motor program, required to perform a sequence of sounds, syllables and words, takes into account earlier the number of remaining phonemes within a breath group. The motor program shortens some, but maintains the consonant constrictions and the recognizability of the stressed syllables. Without a proper programming of the shortening, some phones might be spoken too quickly to be correctly understood. The $short{-}term\ memory$ (STM) theory was proposed by Lindblom, Lyberg and Holmgrem in [84]. The model suggests that a short-term storage of the instructions for speech articulator movements occurs. The storage continuously changes its contents as the instruction leave the queue to implement the actual movements, and new instructions are inserted in the queue for later implementation. The storage capacity is limited, and therefore it has to be economically used. In addition to the right-Ieft coarticulation, sometimes also a $left{-}right$ coarticulation is observed. If an active forward looking planning process is involved in the right-Ieft coarticulation, the left-right coarticulation is the consequence of articulator momentum or low-level movement constraints, and not by a higher-level

motor control program. Formant transitions in vowels, following a preceding consonant, are a good example of this [85].

Coarticulation may cause changes in acoustic speech patterns, produced by articulatory motion, or a change in duration of the phone targets. Monosyllabic words, usually, are shortened more in anticipation of an increase in the number of later syllables than early bisyllabic words. Similarly, trisyllabic words are shortened less than bisyllabic words. According to the STM model, in the production queue storage is expressed in terms of number of syllables, and less of that limited capacity is needed for a word with fewer syllables. The readjustment performed to economize the available storage space is continuously dependent on the number of phonemes, syllables, words or phrases in a sentence.

Chapter 2

Emotion and Mood

2.1 Emotion vs. Mood

Voice is a very powerful communication medium: it allows to convey different and sometimes contradictory meanings at the same time. Non-verbal aspects of speech can play an important role in vocal communication. Message, speaker, language, mood and emotion can influence this very complex signal called speech. Emotional speaking can make one perceive the same textual message in different ways, inducing one to perceive different meanings.

According to Eckman [86], moods usually are emotional feelings lasting for an extended period of time. On the contrary emotions are temporary feelings that tend to come and go quite quickly. If emotions are generally more varied, moods are generally felt in a more generalized way: good mood or bad mood. Moods activate specific emotions. Scherer stated moods might emerge without apparent cause, showing a low intensity, little response synchronization, but a longer duration with respect to emotions. On the contrary he defined emotion as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism" [87].

2.2 Emotions

Generally, emotions produce pervasive and short-lived changes in the organism, representing the answer to stimuli of major significance elicited from one individual. Emotions produce some specific action, mobilizing all the possible resource readiness to face the

respective situation, while providing a latency period that allow adaptation of the behavioural reactions to the situational demands [88, 89]. Such latency period might be used in socially living species to predict the likely reaction of others to an action that is "ready" for execution as the result of a particular emotional state [90]. In addition, as demonstrated by Darwin in his classic work on the expression of emotion, the emotional expression is important to supply the vital function of externalizing an individual's reaction and action propensity and of communication this information to the social environment [91].

Even though, many studies have been devoted to facial expression, vocal communication of emotion can be very effective in nature [92, 93].

Several models have been developed to explain the nature of the expressive communication [94]. According to Ekman et al. [95], some social "display rules" determine partially the human vocalization. Scherer et al. in [96, 97] discussed the distinction between *push* and *pull* effects to determine factors that operate on vocalization. In this model, the physiological changes, that accompany emotional arousal and that consequently modify the voice production mechanism, produce the push effects. On the other hand, pull effects do not depend on the internal physiological processes in the organism, but they originate in external factors, such as ritualized or conventionalized acoustic signal patterns that are required to ensure information transfer. Often, both types of effects determine the acoustic nature of a vocalization. Hence, both emotion-related physiological changes internal to the organism and external constraints or social target patterns are involved, but the understanding and the disentangling of these two kind of effects is not so easy. In fact, many studies do not differentiate between them [90].

To characterize emotions, it is mandatory to have available a suitable emotional speech database [98–100]. For this purpose, it is fundamental being able to evaluate the quality of such databases [101]. Different objectives and methods can be considered when a speech corpus is recorded. Usually three different kinds of speech corpora can be used:

- Actor (Simulated) based emotional speech database
- Elicited (Induced) emotional speech database
- Natural emotional speech database.

Experienced and trained theatre or radio artists are used to collect simulated emotional speech corpora. They are asked to express different sentences in different emotions. Usually these sentences are linguistically neutral. Usually, to take into account the variability due to the degree of expressiveness and physical speech production mechanism,

the recordings are obtained in different sessions. Such a kind of database is maybe preferred because of the easy and reliable method of collecting expressive speech. They can comprise a wide range of emotions. More than 60% of the emotional databases collected are of this kind [100]. The simulated emotions are fully developed or exaggerated. The simulated emotion are typically intense and usually incorporate most of the relevant aspects of the expression of the emotion [102]. Generally, these acted emotions are therefore considered more expressive than real ones [101, 103].

Elicited emotional speech database are formed by artificially eliciting emotions in people, without knowing the real felt emotion. Generally, subjects are asked to involve an emotional conversation with an anchor, who is in charge of creating different contextual situations to elicit, by means of a conversation, different emotions in the subject enrolled. Such a kind of database is definitely more natural, with respect to the simulated one, but they still might be not properly expressive, since subjects still know that they are recorded. In addition, sometimes, subjects can be asked to have such emotional conversation with a computer, whose responses are controlled by a human being without that the recorded subjects know it [104].

Natural emotions can be difficult to record and clearly recognize. They are also known as *underlying emotions*. This kind of emotion can be naturally recorded from call centre conversations, from cockpit recordings during abnormal conditions, during a clinical visit in a dialogue between patient and clinician, during emotional conversation in public places and so on [100]. Anyway it is not so easy to record a wide range of emotions in this way. Moreover, the labelling phase of these emotions can be highly subjective and therefore categorization can be debatable. From the legal point of view, some issue related to privacy and copyright must be taken into account [101, 104].

While emotional elicitation can sometimes have some ethical contraindications regarding the chosen way to induce the wanted emotion, the other two kinds of datasets have to be used carefully. Since the emotions they collect were obtained in a different way, they might be characterized in a different way. Hence the natural emotional speech database should be preferred to the other, despite the difficulty in obtaining the natural emotions and labelling them. For this aim, recently, a series of studies were conducted on the investigation of the differences between simulated or acted emotions and the natural or spontaneous ones. Vogt and André [105] showed that partially overlapping features sets can be used to recognize different emotions from acted and spontaneous speeches. Bänziger and Scherer [106] defended in a detailed way the cautious use of speech materials in the study of acted emotions. In fact, the difficulties when recording different and often rare natural emotional states from the same subjects, and assessing natural emotional states may explain why acted speeches datasets provide an important

contribution to the field. Schuller et al. [107] sustained that acted corpora have two disadvantages: the first is that acting emotions is different from producing "spontaneous" emotions [108], actors mimic the way that people externalize emotions, and secondly, the prompted types of emotions are not the same as those in realistic scenarios. So while the acquisition of realistic corpora is envisaged, using acted corpora could be convenient for benchmarking, even if the relationship between the results obtained from the two kind of datasets is unclear [109].

2.2.1 Prosodic features in emotion

Duration, intonation and intensity patterns are usually imposed by people when they produce speech. The naturalness of human speech is provided by such prosodic constraints. Since prosody is associated with larger units such as syllables, words, phrases or sentences, prosodic features are often seen as a supra-segmental. Prosodic features can convey human emotional expressiveness [100]. Four main levels of prosodic manifestation are recognized [110]. The first is the linguistic intentional level, then the articulatory level, the acoustic realization and finally the perceptual level.

According to several studies reported in the literature, energy, duration, pitch and their derivatives are considered as correlates of the expression of emotions [111–114]. Important prosodic features of emotions are: minimum, maximum, mean, variance, range and standard deviation of signal energy, and of F0 [115, 116]. An attempt to estimate the steepness of the F0 contour during rise and falls, articulation rate, and number and duration of pauses was performed to characterize emotions by Cahn [117] and Murray and Arnott [116]. In the latter, prosodic features were extracted at both syllables and consonant and vowel levels. Murray et al. [118] and Scherer [119] discussed the importance of prosodic contour trends to characterize emotions. The identification of four emotions, i.e. fear, anger, sadness, and joy, was proposed to be carried out by means of the study of the peaks and troughs of the fundamental frequency profile, the intensity, and the durations of pauses and bursts. An average emotion recognition equal to 55% was reported [120]. Minimum, maximum and median values of F0 and slopes of F0 contours are considered salient features since they enable an emotion recognition accuracy of about 80% with a K-nearest neighbour classifier [111]. Iida et al in [121] investigated the complex relations between F0, duration and signal energy features to detect emotions from speech. Luengo et al. identified four emotions in Basque language obtaining a emotion recognition performance of about 92% by using GMMs on acted records[122]. Such a result is achieved by the authors after having extracted 86 prosodic features and having selected the best 6 ones. Kao and Lee obtained similar performances (92% of emotion recognition of four emotions) in Mandarin by using F0 and power-based

features extracted at the frame, syllable, and word levels [123]. Wang et al. classified six emotions from Mandarin language obtaining an 88% emotion recognition rate using SVM and genetic algorithms [124].

According to the literature, most speech emotion recognition studies take into account static (global) prosodic features extracted at the utterance level [111, 113, 114, 121, 125, 126]. The dynamic behavior of prosodic patterns (local) have been explored in few studies [120, 127]. Rao et al. [128] performed an elementary prosodic analysis at speech and syllable levels by using only the first order statistics of the prosodic parameters. The contribution of studying the static and dynamic, and thus global and local, prosodic features extracted at sentence, word and syllable levels may be very important [100]. Recognizing emotions by using shorter speech segments might be useful for the development of a real time emotion recognition.

2.2.2 Source features in emotion

The suppression of vocal tract characteristics, by means of an inverse filter based on linear prediction coefficients (LPCs), allows to estimate from a speech signal the *linear prediction residual* (LPR) that contains information about the excitation source [129].

Several studies of excitations source features demonstrated that they can convey all aspects of speech such as message, speaker, language and emotional cues, even if this kind of features may not compete with other well-known spectral and prosodic features [100]. LPR derived features have been used successfully to extract information regarding pitch for speaker recognition in [130], regarding signal energy for vowel and speaker recognition in [131], and in several other studies.

In the literature, source cues are not exhaustively and systematically explored for speech emotion recognition [100]. This kind of features might provide information regarding the specific emotion, but in the form of higher order relations among linear prediction (LP) residual samples, parameters of instants of significant excitation, parameters of glottal pulse shape and so on [100]. Iliev and Scordilis [132] reported that glottal symmetry can be a simple but quite effective speech feature enabling a high classification performance for spoken emotions. Chauhan et al. [133] reported that the emotion-specific excitation source information might be present in the higher order relations among the samples of the LP residual. In this work, eight emotions were studied and an average emotion recognition performance of about 56% was achieved by means of two models: auto-associative neural network (AANN) and Gaussian mixture models (GMM). Epoch parameters were explored by Koolagudi et al. [134] to recognize the emotion using speech utterances, showing performances of classification above chance level, and thus

indicating the presence of useful emotion specific information. The excitation source was investigated also in [135] by Al-Talabani et al, noticing that the high dimensionality feature space in emotion recognition is a serious challenge. In [136] Gangamohan et al. investigated two features, i.e. strength of excitation and spectral band energy ratio, which are both related to the excitation source component of speech, to discriminate "angry" and "happy" emotions from the speech corpus named Berlin EMO-DB [137]. In [138] Yadav and Kumari investigated the LP residual at a sub-segmental level, segmental level, and supra-segmental level, reporting 58.4%, 65.6% and 48% respectively average emotion recognition rates.

2.2.3 Vocal tract features in emotion

Speech segment lasting $20-30\ ms$ are usually used to extract vocal tract features [100] in the frequency domain. Some features of them are: formants, their bandwidths, spectral energy and slope. The inverse transform of the log magnitude spectrum of a speech frame is known as the cepstrum [14]. Some common features derived from the cepstrum representing vocal tract information are the MFCCs (Mel frequency cepstral coefficients) and the LPCCs (Linear prediction cepstral coefficients). According to Ververidis and Kotropoulos [98] MFCCs, LPCCs, perceptual linear prediction coefficients (PLPCs), and formant features are common features used for emotion recognition. Generally these kinds of features are considered to be strong correlates of the shape of the vocal tract and of articulatory movements [139]

Combining MFCCs, LPCCs, RASTA PLP coefficients and log frequency power coefficients (LFPCs), Pao et al. were able to to classify anger, boredom, happy, neutral and sad emotions in Mandarin [140, 141]. Williams and Stevens [142] represented emotion specific information by means of log frequency power coefficients (LFPC). For this, a four stage ergodic Hidden Markov Model (HMM) was used to classify emotions. LFPC parameters showed comparable performances as LPCC and MFCC features, while LFPCs performed slightly better [142, 143]. F0 variability is supposed to be modelled by MFCC feature extracted from lower frequency components ($20-300\ Hz$) of the speech signal. In this case, they are known as MFCC-low features and were used to perform emotion recognition in Swedish and English emotional speech databases. Neiberg et al. reported that MFCC-low features outperform F0 features in emotion recognition tasks [144]. Speaker-independent emotion recognition was performed by estimating the mel-frequency cepstral coefficients over three phoneme classes: i.e. stressed vowels, unstressed vowels and consonants. They are known as class-level spectral features. The accuracies of their classification were consistently higher than those obtained from prosodic or utterance-level spectral features. Moreover, the combination of class-level

and prosodic features allowed to improve performances. Consonant regions, more than either stressed or unstressed vowels, seem to contribute more regarding specific emotions. Another important result, reported into the study [145] performed by Bitouk et al., is that the average emotion recognition rate and the length of the utterance are proportional. Sigmund [146] carried out both Fourier and Chirp transforms of vowel segments, and reported that the higher frequency regions of speech are appropriate for characterizing stressed speech. The amount of emotion specific information can be conveyed by different portions of the utterance, depending on the emotional expression pattern [134]. Since emotions may be low-grade, their expression in a spectrum may be gradual.

2.2.4 Combination of features in emotion

Recently, some researchers focussed on combining different features to improve speech emotion recognition. The categories of speech features that are discussed in the previous sections can be considered to be complementary. A proper combination of these features might improve overall performances. Gobl and Chasaide [147] highlighted the role of voice quality in conveying emotions by means of spectral and prosodic features. The authors reported that voice quality such as harsh voice, tense voice, modal voice, breathy voice, whisper, creaky voice and lax-creaky voice are more efficient in detecting underlying (mild) emotions than the full blown emotions. No one-to-one mapping between voice quality and emotions was reported, but sets of emotions were associated with the same voice quality [147]. Kwon et al. explored F0 information, log energy, formants, mel based energy, MFCCs with their velocity and acceleration coefficients to classify emotions [148]. Prosodic, mel-frequency cepstral coefficients (MFCCs) and formant frequency features were used to distinguish six emotions in a language, speaker, and context independent way in [149]. Anger, disgust, fear, joy, neutral, sadness, surprise, and teasing emotions collected from 50 male and 50 female native Japanese subjects were discriminated by means of both prosodic (energy and pitch) and spectral features (12 LPCCs) by Nicholson et al. in [150]. Around 50% of recognition rate was reported using neural network classifiers [150]. The identification of emotions in Mandarin language was obtained by Zhou et al. by combining articulatory and spectral features [151]. Long-term spectro-temporal speech features proposed in Wu et al. [152] outperformed the short-term spectral features and prosodic features in recognizing seven emotions of the Berlin emotional speech corpus (Emo-DB) [137].

The combination of long term spectro-temporal and prosodic features results in an average emotion recognition of 86.6% involving seven discrete emotions [152]. Schuller [153] proposed a new method to combine acoustic features with linguistic information to recognize seven discrete emotional states. Emotional phrases are detected from the spoken

word by means of belief networks. Soft decision fusion and neural network classifiers combine acoustic and linguistic information. Acoustic, linguistic and combined information reported respectively an emotion recognition rate of 26%, 40% and 58% [153]. Lee and Narayanan [112] proposed to combine language and discourse information to improve the discrimination between positive and negative emotion for call centre applications. Zhou at al. combined Teager energy values and MFCC features to distinguish between neutral from stressed speech [154].

2.3 Mood disorders

2.3.1 Bipolar disease

Bipolar Disorder is a chronic psychiatric condition [155, 156] that is considered one of the most common and dangerous disorders of affectivity (Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) [157]).

People suffering from Bipolar disease manifest extremely altered mood regulation. They usually experience unbalanced mood swings among *depression*, *mania* or *hypomania*, and *mixed states* (that is a state in which both symptoms of depression and hypomania are present at the same time). Such swings can be cyclic and sometimes extreme. Normally, these mood changes can have a significant impact on the patients' social, occupational, and general functioning and wellbeing. The patients' quality of life can be severely affected by such swings, even during the time periods in which they are free of clinical relevant symptomatology [158, 159]. Such reduction in quality of life can also be related to a significant loss of cognitive performance [160]. Other symptoms such as somatic pain or functional symptoms (headache, dyspepsia, etc.) can be also frequent. Moreover, this kind of patients often experience anxiety, associated with suicide attempts, lifetime alcohol abuse, and psychosis [161]. Suicide, in fact, may occur in up to 20% of the cases [162].

Depression is a very low mood state characterized by sadness and hopelessness. Depressed patients might also experience thoughts of ruin, guilt or death. Differently, mania is a state of hyperarousal that leads to euphoria or irritability, excessive energy, hyperactivity, hypertrophic self-esteem, and a reduction of the need of sleep. Maniacal patients usually express an increased activity and accelerated thoughts, but rather than being a positive condition, these affects cause attention loss and prevent the patient from expressing a coherent mental stream of thoughts. Hyperactivity is often not finalized and patients switching from task to task are not able to complete any activity. Instead, hypomania is the moderate form of mania. Finally, the euthymic state is a

period in which patients do not show enough pathological signs to be considered in one of the above-mentioned clinical states. In the mixed state, since patients share symptoms of both mania and depression, they can be hyperactive but have insomnia, have an increased self-esteem but also thoughts of inadequacy, and so on. Such a disease can therefore be associated with frequent devastating personal, social and vocational consequences.

Bipolar disorder is one of the leading causes of disability worldwide [163]. According to some epidemiological studies performed in United States, almost 15% of the US population has suffered from at least one episode of mood alteration [164], and more than two million Americans have been diagnosed with bipolar disorder. In the United States it was estimated an annual cost of $45 billion [165]. Moreover, it has been estimated that about 27% (equals 82.7 million; 95% confidence interval: 78.5–87.1) of the adult European population, from 18 to 65 years of age, is or has been affected by at least one mental disorder [166, 167]. Despite high managing costs and the high severity of this disease [164, 166–168], in current clinical practice the diagnosis of bipolar disease relies only on interviews and scores from psychological questionnaires, on the physician's own expertise, and on the patient's subjective description of the symptoms. Another important characteristics involved with such pathology is the *comorbidity*, i.e., the simultaneous presence of symptoms which are shared with other psychiatric disorders. All the mentioned issues associated with bipolar disease may lead to subjective interpretations, inconsistencies, and misdiagnoses [169].

Two mathematical models of bipolar disorder were described in terms of low-dimensional limit cycle oscillators by Daugherty et al. in [170]. In their study, rather than focusing on the difficult medical problem of diagnosis, they took into account the dynamics of the models under a proposed treatment strategy, to provide some insight into the complicated dynamics of this disease. Although they stressed the importance of recording longitudinally bipolar patients, they did not use clinical data.

Some authors applied the non-linear measurements to both mood data generally [171] and to mood in bipolar disorder [172–174].

Early studies on speech cue in patients suffering from bipolar disease reported a elongated speech pause time during the depressed phase [175, 176], while phonation time was reported to not vary significantly [176].

Only recently, studies have analysed speech cues in patients suffering from bipolar disease more deeply (Table 2.1). In this framework, mobile devices, as smartphones, have been investigated as aids to the assessment of mental disorder [177, 180, 181]. Both statistics on the use of the smartphone and acoustic features have been investigated [177, 178]. A

TABLE 2.1: Some corpora, tasks and investigated features from the literature for bipolar patients.

Reference	**Corpus**	**Task**	**Feature**
Greden and Carroll (1980) [175]	4 unipolar, 3 bipolar	counting 1-10	speech pause time
Greden et al. (1981) [176]	24 unipolar, 12 bipolar	unstructured interview	speech pause time, phonation time
Muaremi et al. (2014) [177]	6 bipolar patients	phone call	phone call statistics, social cues, acoustic features
Grünerbl et al. (2015) [178]	6 bipolar patients	phone call	phone call statistics, speech features, voice features
Karam et al. (2014) [179]	6 longitudinally collected from bipolar patients	phone call	low- and segment- level features speeches

system aiming at analysing prosodic features in running speech was proposed in [181]. In this study, the authors evaluated the performance of the system, and proposed a case study on a bipolar patient.

In Muaremi et al. [177] speech cues extracted from phone calls were used to assess bipolar disease episodes. In this study, daily phone calls were investigated to carry out the assessment in a real-life environment. Three different kinds of features were extracted: namely phone call statistics, social signals obtained from the phone call conversation and acoustic emotional properties of the voice. Acoustic features showed best performance in terms of state recognition followed by the social cues. All the features together enabled detecting the states with an average score of 83%. Moreover, the authors reported the importance to set individually for each person the features, since each patient behaved differently from the others. Speaking length and phone call length, the harmonics to noise ratio (HNR) value, the number of short turns/utterances and F0 were labelled as the most important variables over all subjects.

Unstructured speech acquired continuously and unobtrusively via the recording of day-to-day cellular phone conversations were performed by Karam et al. [179]. The authors investigated both low-level features, i.e. F0, RMS energy, zero-crossing rate, maximum and minimum value of the amplitude of the speech waveform, and segment level features by using voice activity detection (VAD) measurements, and suggested that manic and depressive mood states can be recognized from speech data.

A system, based on smartphone-sensing, aiming at the recognition of depressive and manic states and the detection of state changes in patients suffering from bipolar disorder was studied in [178] by Grünerbl et al.. Phone call features, i.e. statistics on the phone calls, speech features, i.e. statistics on the verbal interaction of the patients and the other talker, and voice features, extracted with the open-source "openSmile" toolbox [182] and providing different low-level descriptors, were investigated. The average recognition was equal to 76%. An early detection of changes in a patient's state of about 97% was achieved, allowing the authors to state that reliably early warnings can be provided.

Though some efforts have been made in this area, the greatest number of research in the relation between mental disease and voice have been focused on depression.

2.3.2 Depression

One of the most common mood disorder is *clinical depression*, which may be caused by the difficulties in coping with stressful life events. It may cause persistent feelings of sadness, negativity and difficulty in coping with daily responsibilities. In 2002, unipolar depression was listed by the World Health Organisation as the fourth most significant cause of disability world wide, and predicted that it will be the second one by 2030 [183]. In 2010 the cost of depression per patient in Europe was estimated to be equal to €24000, in terms of relative value assessed across 30 European countries, while the total cost of depression in the European Union was estimated to be €92 billion, with €54 billion of this cost derived from the lost work productivity [184]. In addition, in 2002, the cost derived from the lost work productivity in United States was estimated to be $44 billion, corresponding to a difference of $31 billion if compared to the work productivity lost in workers not suffering from depression. The World Health Organisation estimated, in 2014, that every year over 800000 people die from suicide, and, moreover, that at least 20 times more people exist who attempted suicide [185]. Although suicide is a private act, it has a profound negative impact on lives of those who knew the person. According to [186] at least 6 other people are intimately affected by the negative impact of a suicide. Depression often increases the individual's risk to engage in suicidal behaviours [187, 188]. About one person out of two who commit suicide meets the criteria for a clinical diagnosis of a depressive illness [189, 190]. In such a context, it is clear how an effective diagnosis and treatment of depression might play a role in suicide prevention [191]. It might be useful undertaking screening for risk of suicide for all individuals undergoing assessment or treatment for depression [192].

Since a single clinical characterization of depressed individuals is lacking, the diagnosis of depression is subjective in nature and time consuming. Usually gold-standard diagnostic and assessment tools for depression are involved to the opinion of individual clinicians and thus can be subjective. The Hamilton Rating Scale for Depression [193] is one of the more common diagnostic tools and it is based on a interview style assessment. The tool assigns a patient a score that relates to his/her level of depression. Performing a proper diagnosis by means of these tools is not easy. They rely heavily on the ability, desire and honesty of a patient to communicate their symptoms, moods or cognitions when, by definition, their outlook and motivation are impaired [194]. It therefore requires time to perform a proper diagnosis, and large degree of clinical training, practice and certification to produce acceptable results [195]. On the other hand, an objective measure, that could

also be clinically useful for depression is lacking. An objective screening mechanism, based on biological, physiological and behavioural signals could be very useful to enhance the current diagnosis. Several biological markers such as low serotonin levels [196, 197], neurotransmitter dysfunction [198, 199] and genetic abnormalities [200, 201] were related to depression, but up to now no specific biomarker has been found. Although, biomarkers are still lacking, recent advances have been achieved regarding affective computing and social signal processing to develop diagnostic tools for depressive patients [202–206]. Facial and body tracking might detect characteristic behavioural changes relating to depression.

Recently, the automatic detection of mental illness has been often addressed by means of the study of speech, and specifically nonverbal paralinguistic cues have became popular. In fact, since speech can be analysed cheaply, remotely, non-invasively and non-intrusively, it may be a good candidate to be used in an automated system. In addition, clinicians often analyse subjectively the verbal behavior of a patient during diagnosis. Depression is often related to a decreased verbal productivity, a diminished prosody and monotonous and "lifeless" sounding speech [207, 208]. Furthermore, speech quality has been observed to change to a hollow and toneless sound in a person who is becoming suicidal [209]. The speech processing could have a great impact in primary health care settings. It has been estimated that between 50% and 70% of people experiencing depression consult their primary health care provider [210], while General Practitioners have only a 50% success rate when diagnosing depressed people [211]. Therefore, using methods and tools for the early diagnosis could be fundamental regarding suicide prevention, since in up to 66% of suicides, the patients have contacted their primary health care provider within a month prior to their death [191].

Several studies attempted to correlate prosodic, articulatory, and acoustic features of speech to clinical ratings of both depression [195, 206, 212–218] and suicidality [209, 219, 220] as well as researched the development of automatic analysis of speech with a view to the early diagnosis of such illnesses [221–223], but little is known about bipolar disease.

2.3.2.1 Definition of clinical depression: making a diagnosis

Usually clinicians rely on the *Diagnostic and Statistical Manual of Mental Disorders* (DSM), published by the American Psychiatric Association to diagnose mental disorders. Its first edition was published in 1952 and it is in its 5th edition at the moment [18]. The manual was designed to provide standard criteria for the classification of mental disorders, performing a proper classification by means of the observation of the symptoms and clinical course of the disorder.

In the literature there is no a unique theory explaining the causes of depression, but it is generally considered to be a dysfunction, a reduced activity and connectivity, of the cortical-limbic system [224–227]. This dysfunction results from some interactions between environmental factors including stress and emotional trauma, and genetic predisposition [228]. Most people experience some form of depression in their life, but to suffer from the depression as disease, according to the DSM definition, a person has either a depressed mood or a markedly diminished interest or pleasure in combination with four or more symptoms for longer than a two-week period. The list of these symptoms is reported in Table 2.2.

TABLE 2.2: Symptoms associated with depression [18].

Depressed Mood and/or markedly diminished interest or pleasure
In combination with four of
Psychomotor retardation or agitation
Diminished ability to think/concentrate or Increased indecisiveness
Fatigue or Loss of energy
Insomnia or hypersomnia
Significant weight loss or weight gain
Feelings of worthlessness or Excessive/inappropriate guilt
Recurrent thoughts of death or Recurrent suicidal ideation

The DSM has been criticised because different subsets of symptoms are assigned to the same disease, leaving diagnosis open to subjective biases when a proper patient assessment is not done to achieve a diagnosis [194, 229–234]. It is important to notice that at least four of the DSM symptoms listed in Table 2.2 include opposite manifestations (e.g. insomnia versus hypersomnia). At least 1497 different profiles of depression are compatible with such a definition [235]. Two people, sharing no overlapping symptoms, can receive the same diagnosis according to the DSM [236]. The large variation in depression profiles make its diagnosis difficult and very complex, above all in the attempt of fitting the clinical profile of a depressed person into an objective categorical level, i.e. mild or severe depression [211]. Usually the assessment tools are based on an interview such as the *Hamilton Rating Scale for Depression* (HAMD) [193] or self-assessment scales such as the *Beck Depression Index* (BDI) originally published in 1961 and revised in 1996 [237]. Both these tools evaluate the severity of 21 symptoms observed in depression, to return a score which is related to the level of depression. The greatest difference between the scales is the time people have to spend to complete the questionnaires. HAMD is a clinician-rated questionnaire that can be completed in $20-30$ minutes, while BDI is a self-reported questionnaire that can take about $5-10$ minutes. The HAMD has long been considered as the gold standard assessment tool for depression for both diagnosis and research purposes, thought this status is frequently discussed [238, 239]. The rates of severity of symptoms such as low mood, insomnia, agitation,

anxiety, and weight loss are evaluated in the HAMD by the clinicians to give a patient a score. Such evaluation consists in choosing a possible response to each question by interviewing the patient and observing his/her symptoms. Each of the 21 questions has about 3 or 5 possible responses which range in severity; scored between 0-2, 0-3 or 0-4 depending on the importance of the symptom they represent. Then, the scores are summed and the total is arranged into 5 categories: Normal (0−7), Mild (8−13), Moderate (14−18), Severe (19−22) and Very Severe (≥23) [194].

2.3.2.2 Cognitive effects on speech production

An association between cognitive impairments and depression was shown to affect an afflicted individual's working memory [240]. The phonological loop plays an important role in the working memory. In fact, this loop helps to control the articulatory system and to store the speech-based information for a few seconds. According to Christopher and MacDonald [241], the phonological loop is affected by depression causing phonation and articulation errors.

Speech planning is affected by a reduction in cognitive ability and consequent working memory impairments [242]. Moreover, such a reduction is able to impair the neuromuscular motor coordination processes and to alter the proprioceptive feedback loop affecting the feedback of articulator positions [243]. In the literature some studies confirm that significant correlations between depression severity and pause related measurements are related to the difficulty of depressed people in choosing words [244, 245].

Disturbances in muscles tension [89] and respiratory rate [246] were seen to depend on variations on the somatic nervous system (SNS) and autonomic nervous system (ANS). Changes in muscle tone have also been observed to be linked with the GABA neurotransmitter [247]. Both prosody and quality of speech can be affected by such changes in muscles tension and control. Vocal fold behavior is influenced by modifications in laryngeal muscle tension, while changes in respiratory muscles affect subglottal pressure.

2.3.2.3 Prosodic and acoustic features in depression

Prosodic features (Table 2.3) characterize speech at the supra-phoneme level. They describe variations in perceived rhythm, stress, and intonation of speech. Speaking rate, pitch (auditory perception of fundamental frequency), loudness and signal energy variations are commonly investigated, but in practice the fundamental frequency and energy are the most often studied prosodic features.

Early studies on speech in depressed people showed how this kind of patients had some speech abnormalities, i.e. a reduced pitch, a lowered pitch range, slower speaking rate and articulation errors. More in detail, Darby and Hollien [248] reported that a modification in perceived pitch, loudness and speaking rate was observable in depressed patients between before and after treatment. Five potential characteristics: reduced speaking intensity, reduced pitch range, slower speech, flat intonation and a lack of linguistic stress were also highlighted in depressed patients by Hollien [249]. F0 contours were perceptually investigated and they were demonstrated to contain information about a wide range of prosodic information such as F0 variability, speech rate and pause time by Nilsonne and Sundberg [250]. Taking into account the monotonous and "lifeless" descriptors of speech spoken by patients suffering from depression, it is not surprising that several studies reported a correlation between both reduced F0 range and average F0 value, and an increasing level of depression severity [195, 216, 251–257]. But, in some studies, no significant correlation between F0 variables and depression was found [194, 244, 245, 258–261]. These conflicting results might be explained by the heterogeneity of the depression symptoms, the fact that F0 reflects both the physical state of the vocal folds and the speaker's affective state, gender, and a lack of standardization in F0 extraction methods [194].

Psychomotor retardation (PMR) is a phenomenon that is characterized by the slowing of thought and reduction of physical movements. Its effects, jointly sometimes with changes in the speaker's affective state, can result in small disturbances of muscles tension in the neuromuscular system of the larynx. Hence, the observed reduction in F0 variability can be explained taking into account these disturbances. In addition, the increased monotony might be the result of PMR reducing laryngeal control and dynamics [259, 262]. Some studies reported increases in aspiration with depressed speech that could be explained by the lack of laryngeal control. Another cause could be the increase in vocal tract tension tightening the vocal folds. This could induce a less variable and more monotonous sounding speech [208, 255, 257, 258, 263]. Still, this framework is lacking because an increase in muscle tension may induce a decrease in F0 variation, but it is not able to explain the observed decreased average F0. On the contrary, an increase in vocal folds tension should generate an increase in F0. Therefore, one may guess that F0 is not only a vocal fold movement marker, but also a paralinguistic marker, reflecting the expressiveness of speech, and thus affected by many different speaker states and traits [250]. Healthy and depressed persons may show different personality traits and thus variability [264]. In the literature several variables, associated with depression, have been shown to be able to affect F0. For instance: changes to a person's underlying mood [263], level of agitation and anxiety [244, 257], and personality traits [261]. Hence, from comparisons with non-depressed individuals [216, 251, 252, 254], F0 possibly lacks

specificity for depression. Some studies, aiming at the investigation of F0 changes in depression severity over time, have shown very small statistical effects, that are due to large numbers of participants [245].

Signal energy parameters appeared to be problematic also. Reduced variability in loudness due to a lack of speaking effort was reported in patients suffering from depression before treatment. Such deficiency was significantly reduced after treatment [252]. But, only mild correlations of improvements in mean loudness and variation of loudness with patient recovery was detected by Stassen et al. [254]. Moreover depressed patients were found to speak louder than control subjects in Alpert et al. [244], but not at a significant level. A mixed behavior in patients affected by depression was detected by Stassen in [265]. In fact, this study it was either reported a lack of signal energy dynamics and its relative improvement after treatment, or an overly louder speaking before treatment and a decrease to a normal speaking level after treatment. Quatieri and Malyska [259] found a mildly significant negative correlations between variability measurements of signal energy and depression, while a significant positive correlation was found between signal energy rate of change and depression. According to them, at lower levels of depression signal energy rates can be thought as an index of the improvement of motor coordination.

Speech rate is one of the most promising prosodic features for detecting depression [194]. In fact many studied showed that depressed persons speak at a slower rate than controls [175, 176, 248, 249, 260, 266–268]. More recently, Stassen et al. [256] reports that 60% of the enrolled depressed patients showed a speech pause duration that significantly correlated with their HAMD score. Speech pause duration was also reported to be statistically different in depressed patients with respect to healthy control subjects in [244]. Notwithstanding the limited number of enrolled patients, Cannizaro et al. [258] also showed that a reduced speaking rate was significantly correlated with HAMD measurements. Such findings were confirmed by Mundt on larger databases in two studies. In fact, in [195] it was reported that shorter pausing and faster speaking were detected in patients who responded to treatment with a relative decrease of 50% of their HAMD score. In a follow-up study [245], six prosodic timing measurements were found to significantly correlate with depression severity. These are: total speech time, total pause time, percentage pause time, speech pause ratio and speaking rate. Average syllable duration was reported to be significantly longer in depressed persons with respect to the healthy controls ones by Alghowinem et al. [269]. In addition, a positive correlation between average syllable duration and increasing levels of speaker depression was shown in [253]. These two results confirm an overall decrease in speech rate with depression [194]. According to Trevino et al. [218], the speaking rate at the phone level could be possibly more informative. In fact, they reported that extracting phone-specific

features about speech rate, and combining the average phone-duration measurements, that are highly correlated with depression, a stronger relation between speech rate and depression severity was obtained compared to the global measures previously reported. A consistent correlation was also obtained in grouping individual phones by manner of articulation, i.e. vowels or fricatives. The authors highlighted the possible importance of phone-based indicators of speech rate as a biomarker of depression.

On the one hand, speech rate is clearly one of the strongest features able to detect depression, but on the other hand it is still not understood if a decrease in speaking rate is a possible measure of motor retardation or of cognitive impairment [194]. According to Cannizzaro et al. [258], there are two different mechanism to induce a decrease in speaking rate: the first one is motor impairment, while the second one is the insertion of longer or more frequent pauses into an utterance. The latter could be induced by a cognitive impairment if a person has difficulty in choosing the words. In [258], Cannizzaro et al. did not report a significant increase in speech pauses, while speech rate was decreased. Hence these results suggests that motor retardation may induce a decrease in motor speed and agility and therefore slowing speech. In addition, in [244], it is shown how an increase in speech pause measurements did not reflect a decrease in speech intelligibility. Therefore, the authors stated that speech slowing in depressed people is a marker of decreased cognitive functioning related to speaker motivation.

Since there are natural variations in individual speaking, and the clinical profile of depression is quite wide, it is possible to hypothesize that a single feature is not enough to be used as a clinical marker of depression. Nilsonne et al. [255] hypothesized that, since some natural F0 variation is present in healthy subjects, an investigation of F0 variation would be more useful if performed within-patient. In fact, natural variations in speaking result in normalization problems of prosodic features. Moreover Stassen et al. [265] stated that a multivariate approach is required to perform a proper diagnosis of depression from speech features. According to Moore et al. [215] F0 can be considered to be an abstract descriptor of the vocal folds dynamics and therefore it is not able to report information about vocal fold tension. They showed that glottal features may be more useful to detect depression than prosodic features.

2.3.2.4 Source features in depression

The features that are related to the source (Table 2.4) of voice production and to the airflow streaming from the lungs through the glottis are known as source features. Many of them are extracted from length measures of the glottal flow signal [272, 273], even if it is not easy to automatically extract these time instants, thanks to a non-uniform

TABLE 2.3: Some prosodic measures from the literature for low (control) or high levels of speaker depression.

Reference	**Corpus**	**Task**	**Feature**	**Low level of depression or Control**	**High level of depression**	**Significance (Test)**
Nilsonne (1987) [270]	16 depressed persons (both gender) + 16 controls		F0 range (Hz)	21 ± 2	15 ± 2	$p \leq 0.001$ (t-test)
Breznitz (1992) [251]	11 depressed persons (females) + 11 controls	interview	F0 range (Hz)	38.3 ± 11.3	15.8 ± 18.2	$p \leq 0.004$ (t-test)
Alpert et al. (2001) [244]	22 depressed persons + 19 controls	interview (SCID [271])	F0 mean (Hz)	150.6 ± 31.4	142.0 ± 27.2	Not Significant
Mundt et al. (2012) [245]	54 nonresponders + 51 responders to anti-depressive treatment	free speech/ alphabet/ counting/ reading/ sustained vowels	F0 mean (Hz)	153.3 ± 35.7	155.7 ± 33.5	Not Significant
Yang et al. (2013) [261]	10 nonresponder + 16 responders to anti-depressive treatment	Interview	F0 variation (Hz)	0.23 ± 0.1	0.20 ± 0.1	Not Significant
Kuny and Stassen (1993) [254]	30 depressed persons + 30 controls	Counting+reading +counting	Energy per second (mV^2) (associated with a syllable)	11.0 ± 4.8	9.9 ± 3.7	$p \leq 0.01$ (Wilcoxon)
Alpert et al. (2001) [244]	22 depressed persons + 19 controls	interview (SCID [271])	Loudness (dB) (F0 amplitude)	14.2 ± 7.33	18.1 ± 6.37	Not Significant
Alpert et al. (2001) [244]	22 depressed persons + 19 controls	interview (SCID [271])	Mean pause time (s)	0.68 ± 0.136	0.70 ± 0.162	$p \leq 0.05$ (t-test)
Mundt et al. (2012) [245]	54 nonresponders + 51 responders to anti-depressive treatment	free speech/ alphabet/ counting/ reading/ sustained vowels	Total pause time (s)	36.4 ± 19.4	46.3 ± 34.6	$p \leq 0.01$ (t-test)
Mundt et al. (2012) [245]	54 nonresponders + 51 responders to anti-depressive treatment	free speech/ alphabet/ counting/ reading sustained vowels	Pause variability (s)	0.51 ± 0.15	0.61 ± 0.20	$p \leq 0.05$ (t-test)

vocal fold behavior and formant ripple and noise remaining after inverse filtering, which is required to remove the effects of a continually changing vocal tract [274].

Source features also inform about *voice quality*, i.e. the auditory perception of the modification of vocal fold vibration and vocal tract shape. Information regarding phonation types or laryngeal qualities can be obtained by means of features that report irregularity in phonation [147]. Some common voice quality clues are: *jitter*, that is a cycle-to-cycle variability of the glottal pulse length during voicing, *shimmer*, that is a cycle-to-cycle variability of the speech cycle amplitude during voicing, *harmonic−noise ratio* (HNR), that is a ratio between the harmonics and the inharmonics components. It is important to notice that the lack of standardization in extraction methods, i.e. window duration, sampling frequency, and F0 extraction technique, influence both jitter and shimmer estimates, and makes difficult to compare the results obtained from different studies [275, 276]. The kind of vocal task, i.e. sustained vowels or continuous speech, is also a further confounding factor when investigating jitter, shimmer and HNR [277]. The extraction of these cues is easier in sustained vowels, thanks to their intrinsic stationarity, but differences in sound pressure levels both intra- and inter-subjects might produce errors that possibly could make unreliable the comparison [276]. The detection of voiced sections in continuous speech make more difficult the analysis of this audio recordings [277]. In fact, many efforts have been made developing automatic algorithms aiming at the proper segmentation of the speech signal for glottal source analysis [278, 279]

In the literature, not many studies exist on the effect of depression on source cues reporting voice quality. An increased aspiration, that is a cue of air leakage at the glottis, was found in depressed people compared to healthy control subjects [213]. Statistically

significant differences were found by Ozdas et al. in [222] in jitter but not in the spectral slope by means of an F-test. But, a pairwise t-test reported spectral slope and not jitter as significantly different.

Aspiration, which is a perceived excess of airflow estimated by means of a harmonic/noise decomposition technique, jitter, and shimmer were reported to be correlated with depression severity in [259] by Quatieri and Malyska. These results enabled the authors hypothesizing the presence of a motor retardation in depression, that reduces the laryngeal muscle tension resulting in a more open glottis and turbulent airflow. Hönig et al. [253] reported a strong negative correlation between depression and shimmer, spectral harmonics and spectral tilt. Some of these features describe a more breathy phonation in patients suffering from depression. The *Teager Energy Operator* (TEO) features was found to differ statistically significantly between healthy and depressed people [214].

Recently three important studies, authored by Scherer, reported important findings concerning voice quality in depression [205, 217, 280]. In fact, they report in speakers with moderate to severe depression and speakers without depression a statistically significant difference of the *Normalized Amplitude Quotient* (NAQ), i.e. a feature related to the derivative of the glottal flow rate, and the $Quasi{-}Open{-}Quotient$, i.e. a feature related to the amplitude measure of the glottal flow rate. Both these features were estimated by means of the IAIF algorithm [281] in a fully automatic way. Hence, according to these studies, depressed voices can possibly be described by means of a more strained (tense) voice quality, confirming the results reported by Darby et al. [252], Flint et al. [213] and France et al. [221] that showed an increased vocal fold tension in depressed patients.

A boosting of relative energy in the higher frequency range were observed to be related to an increased laryngeal tension and a subglottal pressure [282, 283]. These results may be explained by a potential irregularly in the glottal pulses shape due to an excessive tension and a disturbances in the coordination of the laryngeal musculature [89, 222] under *emotional stress* encountered under depressive and suicidal states [222]. Moreover, Quatieri and Malyska [259] showed a positive correlations of increased high frequencies in the glottal spectrum after a sub-band decomposition with depression.

2.3.2.5 Formant features in depression

Since an increased muscular tension and changes in salivation and mucus secretion are related to variation in speaker's mental state, thanks to the action of the ANS response [89], such phenomena should also be reflected in formant features variations, providing information concerning the acoustic resonances of the vocal tract (Table 2.5).

TABLE 2.4: Some source measures from the literature for low (control) or high levels of speaker depression.

Reference	**Corpus**	**Task**	**Feature**	**Low level of depression or Control**	**High level of depression**	**Significance (Test)**
Flint et al. (1993) [213]	30 depressed persons + 31 controls	reading 4 sentences	Spirantization (present/absent)	0.32 ± 0.43	0.59 ± 0.56	$p \leq 0.02$ (ANOVA)
Ozdas et al. (2004) [222]	10 near-term suicide + 10 controls	recorded treatment sessions	Jitter	0.0165 ± 0.002	0.0217 ± 0.005	$p \leq 0.05$ (t-test)
Ozdas et al. (2004) [222]	10 near-term suicide + 10 controls	recorded treatment sessions	Spectral Slope (kHz/dB)	-83.3 ± 5.46	-75.56 ± 8.53	$p \leq 0.05$ (t-test)
Scherer et al. (2013) [280]	14 depressed persons + 25 controls	virtual human interaction	NAQ	0.098 ± 0.026	0.065 ± 0.035	$p \leq 0.002$ (t-test)
Scherer et al. (2013) [280]	14 depressed persons + 25 controls	virtual human interaction	QOQ	0.360 ± 0.067	0.275 ± 0.096	$p \leq 0.002$ (t-test)

Displaced formant frequencies [195, 213], shown in depressed people, provided evidence for a decrease in articulatory effort with increasing levels of speaker depression [194]. It is possible to hypothesize that these effects can be generated by PMR tightening the vocal tract [213, 221], or a lack of motor coordination that show an opposite behavior with respect of PMR [194]: an improvement of the first corresponds to a decrease of the latter [206, 218, 259, 284, 285]. Another possibly explanation is that it is the result of anti-depressant medication that dry out the vocal tract and mouth, affecting the formant properties and energy distribution [221].

Significant differences between healthy and depressed persons were shown in formant frequencies by Flint et al. [213], specially regarding the second formant location for the diphthong [ai]. The authors, in fact, speculated that the observed reduced F_2 location was generated by a slowing of the tongue in low-back to high-front motion and that this finding was comparable with those obtained from individuals suffering from Parkinson's disease. In this kind of patients, such slowing depends on a depletion of dopamine, and thus they stated that the result of reduced dopamine, the PMR, can induce these similar articulatory errors via the slowing of the articulatory muscles and the increasing of muscle tone. An increased muscular tension could then cause a modification of formant features in depression. In fact, the narrowing of the formant bandwidth could be the result of the increased tension [89]. In addition, increased facial tension and reduced smiling usually shorten the vocal tract, producing the same effects. The tension of the respiratory and the laryngeal muscles are demonstrated to affect phonation, while the tone of the supralaryngeal muscles and the different activation patterns of the facial muscles can affect resonance and radiation characteristics [89].

An increase in formant frequencies (F_1-F_3) and F_1 bandwidth, a decrease in higher formant bandwidths and a relative spectral flattening were reported in depressed persons by France et al. [221]. In [195] Mundt et al. reported a not significant correlation between F_1 and depression, and a mild correlation with F_2 variability. Later, in [245] Mundt et al. reported that both F_1 and F_2 location and variability were not correlated with depression.

In [214], Low et al. reported that the first three formants and bandwidths show statistically significant differences between depressed and control patients. Classification accuracies equal to or higher than 70% were reported by Helfer et al. [285] developing a two class low/high depression classifier by means of features related to the formants dynamics, especially their velocity and acceleration.

TABLE 2.5: Some formants measures from the literature for low (control) or high levels of speaker depression.

Reference	**Corpus**	**Task**	**Feature**	**Low level of depression or Control**	**High level of depression**	**Significance (Test)**
Flint et al. (1993) [213]	30 depressed persons + 31 controls	reading 4 sentences	F2 location (Hz)	1132.7 ± 264.2	944.5 ± 380.8	$p \leq 0.02$ (t-test)
Mundt et al. (2012) [245]	54 nonresponders + 51 responders to anti-depressive treatment	free speech/ alphabet/ counting/ reading/ sustained vowels	F1 location (Hz)	546.8 ± 67.1	558.2 ± 51.8	Not Significant

2.3.2.6 Spectral analysis in depression

Power Spectral Density (PSD) and *Mel Frequency Cepstral Features* (MFCCs) are common spectral features. Similarly to formants, this kind of features has been seen to vary with a speaker's mental status, though there is some disagreement as to the nature of the effect [194]. In some studies a relative shift in energy from lower to higher frequency bands [221, 222] were discovered, while in other ones was reported a reduction in sub-band energy variability [259, 284]. A shift in spectral energy, from about 500 Hz to 500 - 1000 Hz, with increasing depression severity was found, first of all by Tolkmitt et al. in [257], and then by France et al. [221] and Ozdas et al. [222]. On the contrary, such a phenomenon was not reported by Yingthawornsuk et al. [286], where higher energy in 0 - 500 Hz, and lower energy in the 500 - 1000 Hz and 1000 - 1500 Hz bands were reported comparing depressed speech to the remitted speech. The modification of the resonance properties of the vocal tract filter and the open quotients and the skewness of the acoustic source could explain such energy shift as the result of the increase in vocal tract and vocal fold tension [89]. Often the voice produced in these particular settings are described as throaty, strained or tense [89, 287]. An increased vocal tract tension was demonstrated to be associated with a throaty voice quality by means of Magnetic Resonance Imaging (MRI) [287]. Increased muscle tone in the vocal tract tension produced by PMR, observed in depressed speech, result in alterations of spectral properties [221, 222].

To estimate MFCCs, that is one of the most common spectral features, the signal is filtered by means of a bank of non-linearly spaced band pass filters (mel-filters). The selection of their frequencies response is inspired by the cochlear of the human auditory system. An estimation of the spectral contour can be obtained by recording the magnitude spectrum via the filters. High quefrencies are related to the harmonics, and the

first cepstrum coefficient reports the average spectral tilt of the spectrum, for instance. Often MFCCs are combined with *Gaussian Mixture Models* to obtain a popular speech parametrization that has been shown to provide a suitable technique to classify either low/high levels of depression [288, 289] or the presence/absence of depression [212, 269]. In speaker recognition tasks, MFCCs are usually concatenated with time derivatives (delta) features, which convey frame-to-frame temporal information. A significant negative correlation was reported between MFCCs concatenated with time differences, both the cepstral coefficients and average weighted variance, with the degree of depression [288]. Such findings, since they report decreasing temporal variations with an increase in depressive severity, are coherent with the monotonous definition of depressed speech.

A decrease in sub-band energy variability with increasing levels of depression were also reported. Negative correlations of energy variance with depression, though not significant, were reported in [259] by Quatieri and Malyska. In [284], Cummins et al. applied the Log Mean Subtraction (LMS) to sub-band energy coefficients to report spectral variability. They reported a negative correlation between such measurements and the degree of depression. This means that an increasing level of depression is associated with a decrease in the energy variability. They also report that PMR and depression have opposite effects on speech production mechanisms: energy variability and depression are negatively correlated, while energy variability and PMR are positively correlated [284].

The findings of both Quatieri and Malyska [259] and Cummins et al. [284] are consistent with conclusion drawn in Cannizzaro et al. [258] where the reduction of articulation rate was related to increased muscle tone [194].

Recently, the effect of PMR on signal energy and formants were investigated. Increases in phone length were shown with increasing levels of depression in Trevino et al. [218]. Moreover, they found an increased pause length with some depression sub-symptoms, including PMR, stating that such an increase might be depend on the increased muscular tension. Significant positive correlations between signal energy rate and PMR were reported in Quatieri and Malyska [259], and between PMR and sub-band spectral variability in Cummins et al. [284]. This could mean that persons affected by PMR could require a major effort to produce and sustain speech, thanks to the lack of motor coordination [194, 218, 259, 284]. Since depressive symptoms seem to be very heterogeneous, the development of an overall objective marker by means of a symptom-specific speech based approach might be recommended [218, 262].

2.3.2.7 Combination of features in depression

Recently, some efforts have been made to develop systems able to automatically classify speech, aiming at the detection of the presence or absence of depression, or at assessing the severity of depression. It is important to notice that if for the presence/absence of depression the distinct classes are known, regarding severity this is not the case.

Presence of depression

Several sets of features have been investigated to develop automatic speech classifiers to discriminate depressed persons from healthy subjects. The combination of prosodic, voice quality, spectral and glottal features was investigated with this aim by Moore et al. [215], Low et al. [214] and Ooi et al. [290]. Moore et al. [215] reported good classification accuracies and the suitability of the glottal features for solving the presence/absence discrimination problem. In addition, Low et al. [214] showed how both glottal and Teager Energy Operator (TEO) energy features may improve the performances of both single-feature or combined-feature prosodic or spectral based classifiers.

As opposed to Moore et al. [215] and Low et al. [214], who performed a feature space fusion technique, Ooi et al. [290] developed a classification method based on a weighted sum of the intermediate decisions generated by separated GMM classifiers trained on a particular feature. The weighted fusion outperformed the single feature decision, though both glottal and prosodic features showed reasonable prediction capability.

Many studies have focussed on the suitability of single prosodic, voice quality, spectral and glottal features. Some of them, MFCCs and formants and combinations of them reported the strongest performance in detecting the presence of depression by using a GMM [192, 212, 269, 285, 291].

Severity of depression

Researchers, who are developing system aiming at the depression severity identification, also usually investigate features or features set. For instance, Cohn et al. [202] and Trevino et al. [218] studied prosodic variability. In fact, Cohn et al. [202] investigated F0 and the speech/pause ratio as input to a gender independent SVM classifier reporting an accuracy of 79% when classifying patients who were responding or not to the depression treatment. Their work highlighted that timing measures relating to phone length had stronger correlations with depression severity than global measures of speech rate.

A combination of Normalised Amplitude Quotient (NAQ), Quasi-Open-Quotient (QOQ), PeakSlope and Open Quotient Neural Network (OQ_{NN}) with a SVM employing a radial basis kernel [217] was used by Scherer et al. [217] to assess the severity of depression,

reaching an accuracy of 75%. The 9-item Patient Health Questionnaire (PHQ-9) was used to label the patients. In [205] Scherer et al. reported an accuracy of 51.28% by using a combination of NAQ, QOQ and OQ_{NN} and a SVM this time employing a 3rd order polynomial kernel. The authors stated that the unsuitable SVM kernel was the reason of this lower performances, and, in fact, by using linear discriminant analysis they reported an accuracy of 76.92%.

Cummins et al. [284] stated that medium-to-long term spectro-temporal information has strong discriminatory properties when classifying an individual's level of clinical depression from their speech.

The first three formant trajectories and associated dynamic information in combination with Principal Component Analysis (PCA) were used by Helfer et al. [285] to classify high and low HAMD scored patients. Helfer et al. stated that the maximum area under the ROC curve was obtained when both formant trajectories and dynamic information extracted from free response speech or sustained vowels were taken into account during the system training.

In [288] Cummins et al. explored the MFCCs in combination with the GMM-universal background model (UBM) and Maximum A Posteriori (MAP) adaptation to model MFCC data [292]. The authors compared classification accuracies when performing full MAP adaption versus mean-only, variance-only and weight-only adaptation. Strong classification performances were reported of variance-only and weight-only adaptation, but spectral variability was reported to be very important to assess both the presence and the severity of depression.

Chapter 3

Materials and Methods[1]

3.1 Speech corpora

In this thesis, different kinds of speech data were studied. Synthetic and real data were used to evaluate and test the proposed methods. Healthy control subjects were used to investigate the specificity of the proposed features. At the end, an emotional database and a mood database, of audio speech recorded from people affected by bipolar disease, were investigated.

3.1.1 CMU Arctic Database

The CMU Arctic Database [293] provides both audio and electroglottographic (EGG) recordings. The EGG signal is related to the impedance changes during vocal folds contact. Therefore EGG signal processing is an important tool to estimate F0 and F0 variability reliably. In Figure 3.1 one observes, in an example, the relation between these two signals.

In this study the CMU Arctic Database was used to evaluate the reliability and the performances of the proposed methods. In fact, since the here discussed methods aim at processing audio speech signals to estimate F0-related features, the EGG signal provides a useful reference measure. With this aim, the features extracted from audio speech signals and the features extracted from the EGG signals were compared. To estimate F0 and F0 changes from the EGG signal, a 5-coefficient-Daubechies wavelet-based filtering was performed to deprive the signal of low frequency drifts. On the pseudo-periodical detrended signal, a cycle-waveform matching algorithm is used to detect glottal cycle timing for each voiced segment, using a segment-specific average waveform.

[1] Part of this Chapter has been already published in [1–5].

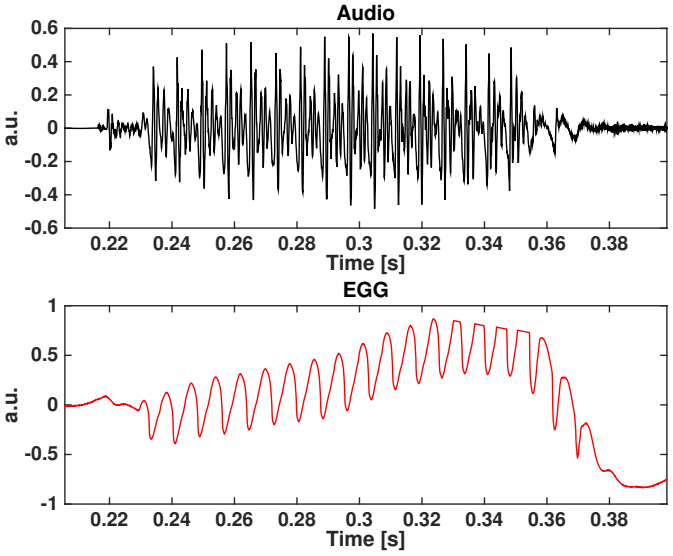

FIGURE 3.1: Example of an audio signal and its corresponding EGG signal.

The corpus consists of approximately 1100 short sentences, corresponding to more than 8000 vowels. Audio and EGG recordings were sampled at a sampling rate of 32 kHz and a resolution equal to 32 bit.

3.1.1.1 Cycle-waveform matching algorithm

The cycle-waveform matching algorithm was applied to the first derivative of the EGG signal (dEGG) (Figure 3.2). First of all, local maxima within every glottal cycle are detected, then the average cycle-waveform shape is estimated by averaging the frames delimited by two consecutive maxima. During this preliminary operation, a average glottal period was estimated as the inverse of the frequency corresponding to the maximum of the spectral amplitude. Finally, by sliding the average cycle-waveform shape along the dEGG signal and computing the estimation of the correlation coefficient (Figure 3.3), the glottal cycle time instants are detected as the ones that showed a high correlation coefficient. The glottal cycle time instants are hence used to estimate the F0-related features.

3.1.2 German Emotional Database

The German Emotional Database [137] is formed by acted speech. Ten different sentences (5 short sentences and 5 long sentences), spoken by ten different actors (5 females and 5 males) simulating four different emotions (anger, boredom, happiness and neutral) were retained.

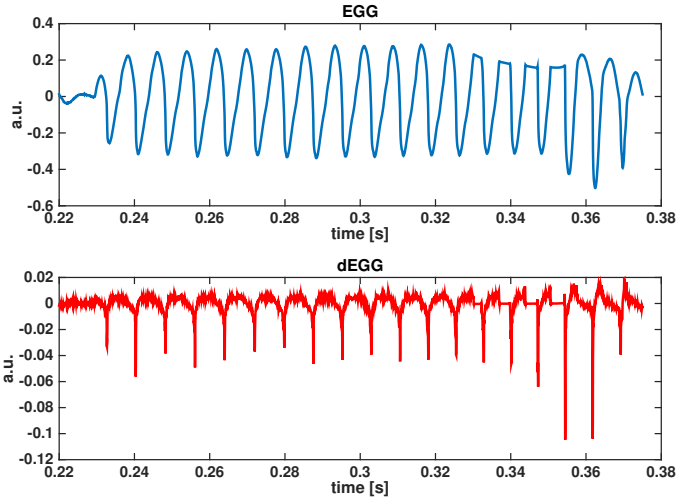

FIGURE 3.2: Detrended EGG (at the top) and its corresponding dEGG (at the bottom).

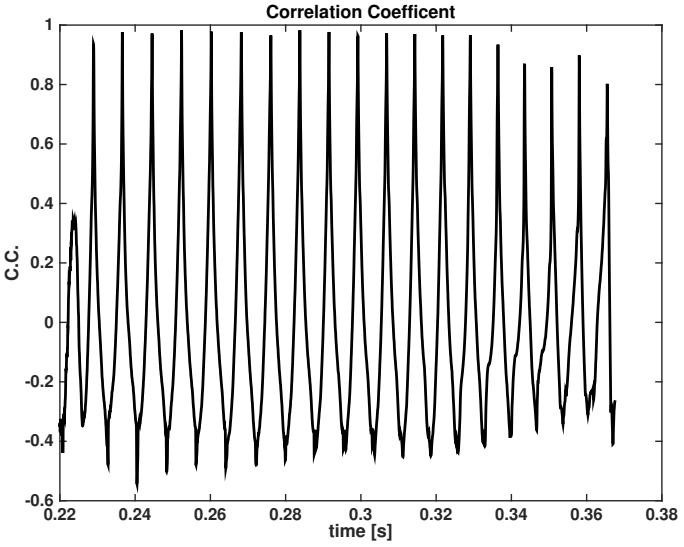

FIGURE 3.3: Correlation coefficients between average cycle-waveform shape and dEGG.

Recordings took place in an anechoic chamber. A Sennheiser MKH 40 P 48 microphone and a Tascam DA-P1 portable DAT recorder were used. Recordings were performed at a sampling frequency of 48 kHz and later downsampled to 16 kHz. A resolution equal to 32 bit float was used. The actors were standing in front of the microphone so they were allowed to use their body language if desired. They were requested to speak in the direction of the microphone with a distance of about 30 cm.

3.1.3 Bipolar Database

In this study, patients suffering from bipolar disease were enrolled within the Psyche European project [294]. PSYCHE (Personalised monitoring SYstems for Care in mental HEalth) project aimed at developing a personal, cost-effective, multi-parametric monitoring system. Such a system was based on textile platforms and portable sensing devices to allow long- and short- term monitoring of the enrolled patients affected by bipolar disorders. These patients were assessed and recorded in two different medical centres: Strasbourg University Clinic, Strasbourg, France and Azienda Ospedaliero Universitaria di Pisa in Pisa, Italy. All subjects had a clinical diagnosis of bipolar disorder, had the competence to lead independent and active lives, and had no substance use disorders. They did not show suicidal tendencies or suffered from delusions or hallucinations.

Seven psychiatric patients (2 females, 40.86 ± 9.56) were recruited in Strasbourg (Table 3.1), while four patients (1 male, 38.5 ± 9.14) were enrolled in Pisa (Table 3.2). Subjects were recorded in two or three different days. Before each session, a physician labelled the patients' mood status by clinician administered rating scales. Four different states were identified: depressed, euthymic, hypomanic and mixed. In each day, the experimental protocol, which received the hospital ethics committee approval, consisted of two sub-sessions, organized as follows:

- TAT (Thematic Apperception Test) images elicitation: the subject had to comment a series of TAT images [295].
- Neutral text reading: subjects read a text that was supposed not to elicit a strong emotional reaction

TABLE 3.1: Patients suffering from bipolar disease enrolled in Strasbourg.

subj.	**gender**	**age**
A	M	40
B	M	53
C	M	40
D	M	28
E	F	34
F	M	54
G	F	37

TABLE 3.2: Patients suffering from bipolar disease enrolled in Pisa.

subj.	**gender**	**age**
H	F	32
I	M	52
L	F	36
M	F	34

The patients enrolled in Strasbourg repeated the "Neutral text reading" twice at each acquisition day. On average, each task lasted from 3 to 5 minutes. Audio signals were recorded with two professional directional microphones (AKG Perception P220 Condenser Microphone). One microphone was used to record the clinician's speech and to allow automatic detection of patient speech. The audio interface was the M-Audio Fast-Track. The sample frequency was equal to 48 KHz and the resolution was 32 $bits$. The microphone was placed on the table at 25 cm from the speaker's mouth.

3.1.4 Healthy Control Subjects Database

18 healthy control subjects (9 females and 9 males, 30 $\pm$ 5 year) were also recruited. Healthy control subjects did not report any actual or past psychiatric disorder, and had no history of neurological or major somatic conditions. At the moment of the study they were not taking any medication. The subjects were recorded according to the same experimental protocol that was used to obtain audio data from bipolar patients, but only ten of them were asked to comment the TAT images. Typically, the second session was recorded 7 days after the first one.

3.1.5 Synthetic data

Two dataset were synthesized to test the Voice Activity Detection (VAD) algorithm and the F0-estimation algorithm, discussed in the following sections. The synthetic vowels were developed by using the parameters reported in Tables 3.4 and 3.5. Vocal tract model was inspired from the one proposed by Klatt and Klatt [296]. In fact, vocal tract is modelled as a cascade of formant and antiformant filters. The coupling between the phonation part and the vocal tract takes into account the modulation of the bandwidth when the glottis opens as well as the tracheal pole-zero pair. The flow rate was modelled by means of the Riccati Ordinary Differential Equations. These equations were solved numerically via a predictor-corrector based on Heun's method. Glottal physiological and neurological tremors were modelled according to the modulation noise model proposed by Steiglitz [297], while muscle jitter was modelled according to the Zolzer's model [298, 299].

A different synthetic dataset was developed to test the F0-corrected LTAS algorithm (see related section). At this aim an autoregressive moving average exogenous ($ARMAX$) model [300] was used to synthesize voice samples. Two different F0 mean values with different applied jitter values were taken into account in this synthesis. The parameters, estimated from a male [a] vowel, were estimated according to model orders for the AR, MA and X parts equal to 16, 4 and 2 respectively.

3.2 Voice activity detection

The detection of voiced segments is required when F0-related features are needed to be investigated. In this study a method of Voice Activity Detection (VAD) dependent on the estimation and the evaluation of the signal intensity across time and of the Zero Crossing Rate (ZCR) [301] was developed. To test the performance of the method, an existing VAD algorithm was used to compare the segmentation results. This method, already reported in the literature in the work of Blanco et al. [302], is based on the estimation and on the evaluation of the autocorrelation function and of the signal energy.

3.2.1 Benchmark method: autocorrelation function and signal energy

The method described in this section [302] and used to test the performances of the proposed one involves two steps. In the first, the signal energy is estimated frame by frame by means of a sliding window. A Hamming window of 30 ms and a time hop of 10 ms were chosen. In this first step, the silent frames were detected and removed. For that, a threshold for silence versus speech activity detection was set to 20% of the global average energy of all the frames of one utterance. In the second step, the autocorrelation coefficient ρ_{ss} between an analysis frame and the analysis frame delayed by 62.5 μs was estimated. According to this estimation, the frames were labelled as voiced if the auto-correlation coefficient was large, and in particular if:

$$\rho_{voiced} = \rho_{ss}(62.5\mu s)/\rho_{ss}(0s) \geq 0.9 \tag{3.1}$$

Only the frames that were characterized by a high relative energy and a high autocorrelation function were labelled as voiced (Figure 3.4).

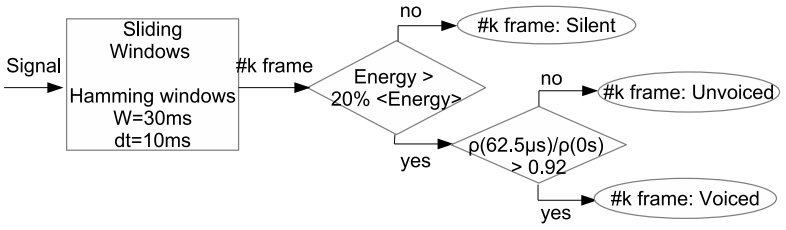

FIGURE 3.4: VAD - Benchmark method: Intervals are labelled as voiced if they report both energy and autocorrelation coefficients rate higher than their respective thresholds.

3.2.2 Proposed method: signal energy and Zero Crossing Rate

Voiced sounds are characterized by high energy values and by lower frequency components with respect to unvoiced sounds. Hence, according to this hypothesis, signal energy across time was estimated using the autocorrelation method as applied to a sliding window [1]. The energy values were obtained by retaining only the frequencies between 5 Hz and 5 kHz. Unvoiced segments were discarded by means of a threshold, related to the median value, applied to the logarithmic transform of energy signal. The threshold level was adjusted according to sensibility and sensitivity criteria. To detect syllables nuclei, local maxima and local dips of the obtained intensity contour were analysed. Using an approach similar to [303], a syllable nucleus was considered to be centred around a local maximum whose intensity was 1 dB higher than the intensity of a preceding local dip. To discriminate between voiced and possible high intensity unvoiced sound, zero crossing rate (ZCR) [301] was estimated according to equation 3.2:

$$ZCR = \sum_{n=0}^{N-2} \frac{1 - sgn[s(n)]sgn[s(n+1)]}{2} \tag{3.2}$$

Only the segments that presented a low ZCR value were considered as voiced (Table 3.3). In Figure 3.5 it is possible to observe a concise scheme of the proposed method.

TABLE 3.3: VAD - Proposed method: Classification of speech signals after energy and zero crossing rate.

ZCR	Energy	Label
Low	High	Voiced
High	Low	Unvoiced

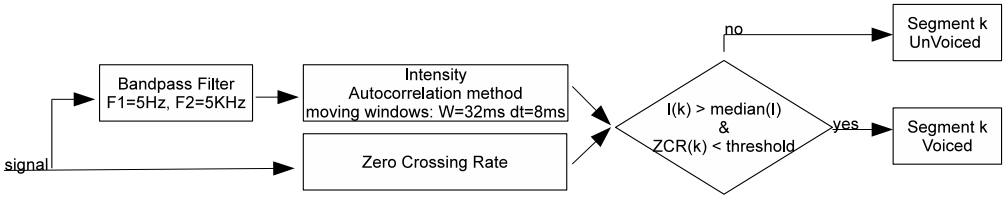

FIGURE 3.5: VAD - Proposed method: Flowchart of the voiced segment detection step. Only the segments having high intensity and low zero crossing rate are considered voiced.

3.2.2.1 Proposed VAD method: Parameter settings

The effects on VAD performance of the sliding window parameters, i.e. width and time hop, as well as of the threshold used to detect voiced segments, were investigated. The CMU Arctic Database [293], which includes audio and electroglottographic recordings,

was used. Window lengths of 16 ms, 32 ms, 48 ms, 62 ms, and 80 ms and time hops of 8 ms, 12 ms and 16 ms were analysed. In addition, threshold levels, used to discard unvoiced segments, ranging from -6 to 6 dB were investigated. The best parameter set was chosen as the one that showed a specificity higher than 0.9 and a sensitivity higher than 0.8, considering the segmentation obtained from the EGG as ground truth. The aim was to minimize the risk of detecting and processing unvoiced segments in place of voiced ones.

3.2.3 VAD methods comparison: Testing on synthetic data

The outputs of the proposed and benchmark methods differ. In fact, the benchmark is able to label the usual three categories: voiced, unvoiced and silence. The proposed VAD is only able to detect voiced frames, separating them from the remaining speech segments, but it can, in addition, detect syllable nuclei.

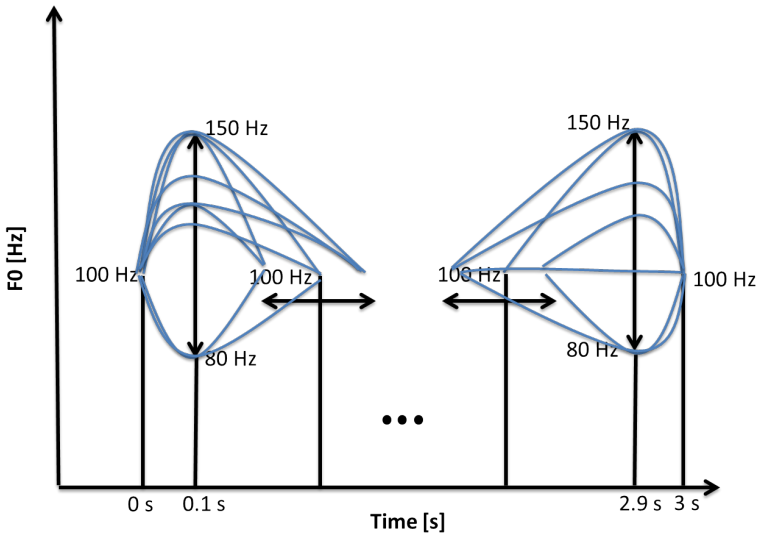

FIGURE 3.6: Scheme of the explored F0 transitions synthesized to test VAD algorithms. In blue some of the explored F0 trajectories are reported.

To test the performance of the proposed method, a study on synthetic audio signals was also carried out. A set of 72 audio signals, each one 3 seconds long, was synthesized. The synthesis produced two voiced segments separated by an unvoiced one. Every voiced segment was designed to explore different F0 contours, which were varied at the end of the first voiced segment and at the beginning of the second, to generate different voiced-unvoiced transients. Each F0 contour, within each voiced segment, started from an initial F0 value equal to 100 Hz and ended in a final F0 value of 100 Hz. Between these two fixed F0 values, each voiced segment explored different F0 trajectories. The parameters were: the F0 targets between the two boundaries, and the target to boundary

to target transient length. Target F0 values ranged from 80 Hz to 150 Hz, while the vowel lengths were 0.6, 0.9, and 1.2 s. Vowel lengths were modified by moving the ending time instants of the first voiced segment, obtaining different target to stop F0 transitions, and by moving the beginning time instants of the second voiced segment, obtaining different start to target F0 transitions. Target F0 values were always reached at 0.1 s (first vowel) and 2.9 s (second vowel), while the starting time instants of the first vowel was set to 0.0 s, and the ending time of the second vowel was set to 3.0 s (Figure 3.6).Table 3.4 reports the parameter values.

Since the explored parameters were the target F0 values in both vowels, and the ending time of the first vowel and the starting time of the second one, F0 contours at the end of the first vowel and at the beginning of the second vowel are expected to vary.

3.3 F0 estimation algorithm

To obtain F0 estimates within each voiced segment an approach based on Camacho's Sawtooth Waveform Inspired Pitch Estimator (SWIPE') algorithm [16] was used. The SWIPE' algorithm measures F0 by estimating average peak to valley distance at harmonic locations. For this aim, a comparison between the spectrum of an audio signal frame and a spectral cosine-kernel was performed. The comparison is obtained by computing a normalized inner product between the spectrum of the signal and spectral cosine-kernel. A weighting of the kernel "spectral-lobes" according to a $1/\sqrt{(f)}$ law was performed. Such a weighting process results in a emphasis of the strongest harmonics with respect to the weakest ones. This choice matches the decay trend of the harmonics of vowels sounds.

The analysis window width is chosen to make the width of the main "spectral-lobes" of the signal spectrum match the width of the positive "spectral-lobes" of the spectral cosine. In the SWIPE' algorithm, as opposed to the SWIPE one, only the first and prime harmonics harmonics of the signal are taken into account, resulting in a significant reduction of the subharmonic errors commonly found in other pitch estimation algorithms [16]. The authors reported that using as many harmonics as possible, up to a certain frequency (usually the Nyquist frequency) outperforms the methods that use a fixed number of harmonics. To avoid that subharmonics of F0 can be estimated as real F0, non-prime "spectral-lobes" related to non-prime harmonics of the signal are removed from the kernel.

The approach, based on Camacho's SWIPE' and used in this study, uses a window size related to the F0 to be estimated. F0 is estimated for each voiced segment by using a

TABLE 3.4: VAD test: parameters used to synthesize audio samples. The symbol "-" indicates that the parameter values vary during synthesis.

Parameter						
F0 tuple (Hz)	100	-	100	100	-	100
F1 tuple (Hz)	710	710	230	230	315	315
F2 tuple (Hz)	1150	1150	2000	2000	605	605
F3 tuple (Hz)	2700	2700	3000	3000	2405	2405
F3 tuple (Hz)	2700	2700	3000	3000	2405	2405
Amplitude tuple (Hz)	0.0	1.0	0.0	0.0	1.0	0.0
Amplitude timing (s)	0.0	0.1	-	-	2.9	3.0
Lung pressure (kPa)	0,5					
Jitter amplitude	0					
Jitter bandwidth (Hz)	200					
Jitter frequency (flutter) (Hz)	50					
Jitter gain (flutter gain) (dB)	0					
Tremor amplitude	0					
Tremor amplitude (wow)	0					
Tremor bandwidth (Hz)	5					
Tremor bandwidth (wow) (Hz)	10					
Tremor frequency in Hz	5					
Tremor frequency (wow) in Hz	2					
Speech sampling frequency (Hz)	50000					
Trajectory sampling frequency	10000					
Area sampling frequency (Hz)	200000					
Number of controllable formants	5					
Relative tract length	1					
Distance between trachea pole / zero	200					
Bandwidth multiplier (open glottis)	2					
Order of FIR filter	51					
FIR filter cut-off	0,08					
Area leakage	0					
Coupling matrix	[[0.0]]					
Open quotient	0,6					
Amplitude quotient	1					
Relative abduction	0,6					
Relative tract length	1					
Eps for max	0,2					
Eps for min	0,5					
Male	TRUE					

sliding window and by using SWIPE' twice (Figure 3.7). $f0$ being a first estimate of F0, the length and the hop of the sliding window, used to obtain the final estimate, are fixed as follows [300].

$$T = 4/f0 \tag{3.3}$$

$$\Delta T = T/4 \tag{3.4}$$

A F0 value is estimated at every step.

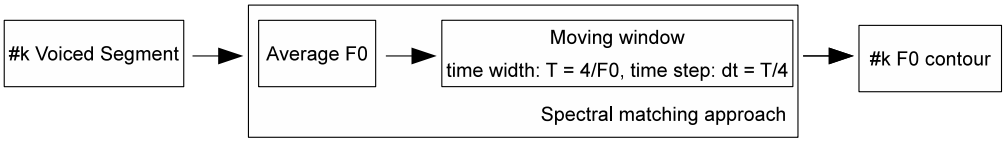

FIGURE 3.7: Flowchart of the F0 contour estimation step. The spectral matching approach was performed by using the Camacho's Swipe' algorithm [16].

3.3.1 F0 estimation algorithm: Testing on synthetic data

To test the proposed method, 512 audio signals were synthesized. All the audio signals lasted 3 seconds, and had five different F0 targets. Two of them, the first and the last were set to 100 Hz and they were located at the beginning (0 s) and at the end (3 s) of the synthetic vowels. The remaining three F0 targets were located at 1.2, 1.5 and 2.4 s respectively and they were modified in the range between 80 and 150 Hz with a step of 10 Hz. The three F0 targets enabled simulating the possible F0 contours: rising, falling, flat, peak and valley. Table 3.5 reports the parameters.

3.4 Vocal features

Hereafter, voiced segments are expected to be interpreted in terms of syllable nuclei because the VAD only reports voiced segments the intensity of which exceeds a threshold. The threshold depended on the median of the frame intensity, and it was properly selected in terms of specificity and sensitivity.

3.4.1 F0, F0 standard deviation, frame-to-frame Jitter

After detecting and segmenting speech signals into voiced segments, and estimating F0 by means of the Camacho's SWIPE' [16], a set of vocal parameters were obtained for each voiced segment [1]. They were the average F0 (meanF0), the standard deviation of F0 (stdF0), and the frame-to-frame jitter (LpJ), calculated in accordance with equation 3.5:

$$LpJ = \frac{1}{N-1}\sum_{i=1}^{N-1}|F_{i+1} - F_i| \Big/ \frac{1}{N}\sum_{i=1}^{N} F_i \tag{3.5}$$

TABLE 3.5: F0 estimation test: parameters used to synthesize audio samples. The symbol "-" indicates that the parameter values vary during synthesis.

Parameter						
Formant timing (s)	0.0	0.75	1.35	1.65	2.25	3.0
F1 tuple (Hz)	710	710	230	230	315	315
F2 tuple (Hz)	1150	1150	2000	2000	605	605
F3 tuple (Hz)	2700	2700	3000	3000	2405	2405
F0 tuple (Hz)	100	-	-		-	100
F0 timing (s)	0.0	1.2	1.5		2.4	3.0
Amplitude tuple	0.0	1.0			1.0	3.0
Amplitude timing (s)	0.0	0.1			2.9	3.0
Lung pressure (kPa)	0,5					
Jitter amplitude	0					
Jitter bandwidth (Hz)	200					
Jitter frequency (flutter) (Hz)	50					
Jitter gain (flutter gain) (dB)	0					
Tremor amplitude	0					
Tremor amplitude (wow)	0					
Tremor bandwidth (Hz)	5					
Tremor bandwidth (wow) (Hz)	10					
Tremor frequency in Hz	5					
Tremor frequency (wow) in Hz	2					
Speech sampling frequency (Hz)	50000					
Trajectory sampling frequency	10000					
Area sampling frequency (Hz)	200000					
Number of controllable formants	5					
Relative tract length	1					
Distance between trachea pole / zero	200					
Bandwidth multiplier (open glottis)	2					
Order of FIR filter	51					
FIR filter cut-off	0,08					
Area leakage	0					
Coupling matrix	[[0.0]]					
Open quotient	0,6					
Amplitude quotient	1					
Relative abduction	0,6					
Relative tract length	1					
Eps for max	0,2					
Eps for min	0,5					
Male	TRUE					

where F_i is the estimated F0 at the $i-th$ frame of each segment. Since the proposed method of F0 takes into account analysis windows containing approximately four vocal cycles, the frame-to-frame jitter is an under-estimate of the actual jitter.

3.4.1.1 Vocal features: Tests on CMU Arctic Database

The performance of the feature extraction, and therefore the whole chain from the detection of the voiced segments to the final features, was evaluated on the CMU Arctic Database [293]. Since the database includes audio and EGG recordings, it can be used to evaluate the reliability of the extracted vocal features.

Since our approach estimates F0 via a sliding window of length $T = 4/f_0$ with a time hop equal to $dt = T/4$, but F0 values were estimated from the EGG cycle-by-cycle, the latter were smoothed in the time domain. In particular, final EGG-derived F0 estimates at the i-th hop, were obtained as the average of four consecutive F0 values. However, unsmoothed jitter (cycle-to-cycle Jitter) is also discussed. Feature comparison is carried out via the correlation coefficient and the slope of the linear model regressing EGG on audio features.

The reliability of the feature extraction was also studied in different noisy conditions. The noise that was added to the audio signals were street noise, train noise, Gaussian noise and echo. The resulting average signal to noise ratio (SNR) is equal to 2.6 dB in the train noise case, and 2.7 dB for the street noise. The Gaussian noise is added to realize SNRs equal to 15, 10 and 5 dB. The noise levels were chosen so as to keep good intelligibility [304]. The time delays, to obtain the echo effect, were equal to 20, 50 and 100 ms.

3.4.1.2 Vocal features: Statistical analyses

The vocal features were used to investigate possible statistically significant differences in thee kinds of database: emotional, bipolar and healthy control subjects databases.

The German Emotional Database [137] was studied to investigate possible statistically significant differences across emotional states. First, normality of the segmental feature distributions was verified by means of a Kolmogorov-Smirnov test. An intra-subject analysis was carried out by means of a Mann-Whitney U test for non-Gaussian segmental features, while for the others a t-test was adopted. These two tests were used to compare the feature distributions, speaker by speaker, between all the couples of emotions. Global average meanF0, global average stdF0, and global average LpJ were estimated for each subject in each emotional state. A one-way ANOVA was used to detect differences among different emotional states in a grouped-subject analysis. The one-way ANOVA was selected because of the normality observed in the distributions of the global average features for each state. Detected differences were considered statistically significant if the corresponding p-values were lower than 0.05.

The proposed features were also used to investigate possible differences in the Bipolar Database and Healthy Control Subjects Database. The average F0 (meanF0), F0 standard deviation (stdF0) and frame-to-frame jitter (LpJ) were obtained for each voiced segment of the audio files of each patients. An intra-subject statistical analysis was performed to investigate changes in the segmental speech features between the sessions. No comparisons were made between features related to different tasks (reading versus free speech). A Kolmogorov-Smirnov test was applied to verify the normality of the segmental feature distributions. A Mann-Whitney U test was adopted for non-Gaussian segmental features, otherwise a t-test was applied. Null hypotheses of equality of means or medians was accepted if the obtained p-value was higher than 0.05.

Inter-state changes were investigated using a non-parametric Friedman's test for paired global average at the group level to discover coherent features changes from mood to mood in bipolar patients and between the two different sessions in healthy control subjects. Anyway it is important to say that the limited number of enrolled bipolar patients did not allow to perform a proper approximation of the Friedman F statistics with the χ^2 distribution. Thus, the estimation of the corresponding significance, i.e. α, was not considered reliable. For this reason, the critical values for Friedman's test reported in [19], obtained through simulations, and corresponding to a statistical significance equal to 0.05, were used. Differently, inter-session analysis on subjects forming the Healthy Control Subject Database was investigated using only a non-parametric Friedman's test for paired data without the need of using the critical values reported in [19]. Also in this case, Friedman's test was applied on global average features.

3.5 Prosodic features

3.5.1 Taylor's Extended Intonational Model: An application to the syllable nucleus

Recently, the interest has grown in models that improve the description of the dynamics of speech features. In particular, the relevance of shape, slope and range of the F0 contour in emotional speech perception, synthesis and automatic recognition has been described [100, 305–307]. Moreover, local features that describe the temporal dynamics have been found to complement global, prosodic features [307] (e.g. intonation contour at the sentence level). Focussing on rising and falling intervals of the stylized fundamental frequency contour [106] highlighted how the contour slope tends to be steeper in higher arousal states. The phenomenon of F0 declination across an utterance was also studied

in emotional speech [308]. Moreover, it was found that the F0 contour slope in the last syllable of an utterance may convey different moods [309].

In this study, an investigation of the description of F0 contours [2, 3] is carried out with a view to discriminating among different emotions, and distinguishing different mood states in bipolar patients. A formal description of F0 contours in syllable nuclei is discussed. Two categories of features are proposed. The first is borrowed from Taylor's Tilt Intonational Model [17] and it describes morphologically F0 contours in voiced segments. Unlike Taylor, the proposed features are extracted for every syllable nucleus and not only for intonational events (i.e. pitch accents and boundary tones). The second category of features is related to the speed of F0 variations and estimates the steepness of both rising and falling F0 contours in each voiced segment.

The following feature are borrowed from Taylor's Tilt Model [17]. They report the "relative sizes of the amplitude and durations of rises and falls" of the contour. Within each voiced segment, the local contour maximum is detected and the features are estimated as follows (equations 3.6, 3.7, and 3.8), starting and stopping at the segment boundaries.

$$Amplitude^* = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{3.6}$$

$$Duration^* = \frac{|D_{rise}| - |D_{fall}|}{|D_{rise}| + |D_{fall}|} \tag{3.7}$$

$$Tilt^* = \frac{Amplitude^* + Duration^*}{2} \tag{3.8}$$

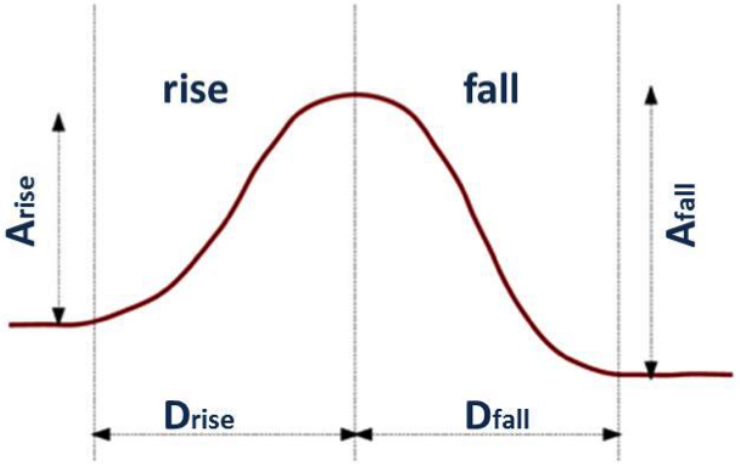

FIGURE 3.8: Parameters of the Tilt Model.

A_{rise} and A_{fall} are the F0 changes during the rising and falling intervals, D_{rise} and D_{fall} are the duration of the rising and falling intervals (see Figure 3.8). Figure 3.9 shows possible F0 contours and their amplitude* feature values.

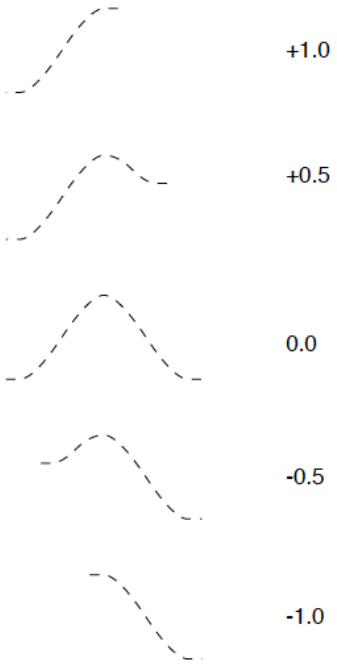

FIGURE 3.9: Examples of 5 contours with their amplitude* values [17].

Amplitude* (ampl*) feature is an index of the difference between the F0 excursion during rising and falling. Duration* (dur*) reports the time intervals over which rising and falling occur. Finally tilt* is the mean value of the amplitude* and duration*.

The previous features describe the shape of F0 contours in voiced syllable nuclei, but they are insensitive to temporal scale. Thus, differently from Taylor, a second category of features that takes into account the speed of F0 change is considered. The steepness of the F0 contour during rising (PosSlope) (equation 3.9) and falling (AbsNegSlope) (equation 3.10) is estimated.

$$PosSlope = |A_{rise}|/|D_{rise}| \tag{3.9}$$

$$AbsNegSlope = |A_{fall}|/|D_{fall}| \tag{3.10}$$

Finally, two other features are estimated according to equations 3.11 and 3.12:

$$SumDer = Slope_{rise} + Slope_{fall} \tag{3.11}$$

$$GlobalSlope = \frac{|A_{rise}| - |A_{fall}|}{|D_{rise}| + |D_{fall}|} \tag{3.12}$$

SumDer (equation 3.11) is a sum absolute slope value. GlobalSlope (equation 3.12) is defined as the F0 slope between the boundaries of each voiced segment.

3.5.1.1 Taylor's Extended Intonational Model: Statistical analysis

The features were estimated on the emotional speech database first [137], then on the bipolar and healthy control subjects databases.

As regards the emotional speech database, statistical tests were performed to evaluate differences among different emotional states. The tests were performed both at single subject level (i.e. intra-subject) and at the group level (i.e. inter-state).

Parametric and non parametric statistical tests were used according to feature distributions. Gaussianity of feature distribution was tested using a Lilliefors test. The non parametric statistical tests were the Mann-Whitney U-test for intra-subjects analysis and the Kruskal-Wallis test for inter-emotion analysis per group. The parametric test employed was the one-way ANOVA for inter-emotion analysis per group. These two inter-emotion analysis per group were performed on global average features.

As regards bipolar patient data, intra-subject analyses were performed to test for statistically significant features changes between mood states. Such an analysis was performed by means of a Mann-Whitney U-test. The comparison was only performed between feature sets reporting the same task.

Inter-state changes were investigated using a non-parametric Friedman's test for paired data at the group level on global average features to discover coherent features changes from mood to mood in bipolar patients. Also in this case the limited number of enrolled bipolar patients did not allow performing a proper approximation of the Friedman F statistics with the χ^2 distribution. For this reason, the critical values for Friedman's test reported in [19], obtained through simulations, and corresponding to a statistical significance equal to 0.05 were used.

To test for specificity of the proposed features with respect to mood changes, features estimated from different recording sessions, but identically labelled were compared by means of a Mann-Whitney U-test. This was accomplished by comparing morning and afternoon recording sessions from bipolar patients.

Moreover, a comparison between the two-days-related-recordings of subjects forming the Healthy Control Subject Database was performed. In this latter case both a Mann-Whitney U-test and a Friedman's test were used. The first one was used to study intra-subject feature changes, while the second was applied to investigate possible coherent feature changes between the two recording sessions. The Friedman's test was used on global average features directly without the need of using the critical values reported in [19].

3.5.2 Spectral analysis of Intonational contours

In the literature, the speech intonation contour has been found to be a reliable indicator of mood changes from a euthymic to an either depressed or manic state [3]. Despite the relevance of the results, several limitations have been observed. Particularly, the direction of the features changes was not coherent across subjects. Better consistency may be achieved both by improving subject status characterization, e.g. by evaluating anxiety level [3], and by investigating other features. In this work a spectral analysis of the F0 contour [4] is proposed to investigate differences in mood states in patients suffering from bipolar disorder.

In a first step, voice activity detection (VAD) is carried out by means of autocorrelation coefficients and speech energy, which reports any voiced segment and not prominent syllable nuclei only. Later, the F0 contour is estimated within any voiced segments by means of Camacho's SWIPE' algorithm [16]. A cubic spline interpolation is used to obtain F0 contours in unvoiced segments, while F0 in silent pauses is set to 0 Hz. Finally a set of 7 features is extracted from the spectrum of each mean-subtracted F0-contour. Power spectral density is estimated from each recording using the periodogram. The features are: median frequency (F_{median}), power amplitude at the median frequency (A_{median}), maximum peak power amplitude (A_{peak}), and the corresponding frequency (F_{peak}), the ratios between amplitudes and corresponding frequencies, and *Slope* according to (equations 3.13, 3.14, 3.15) (Figure 3.10).

$$Ratio_{peak} = A_{peak}/F_{peak} \tag{3.13}$$

$$Ratio_{median} = A_{median}/F_{median} \tag{3.14}$$

$$Slope = \frac{A_{peak} - A_{median}}{A_{peak} - A_{median}} \tag{3.15}$$

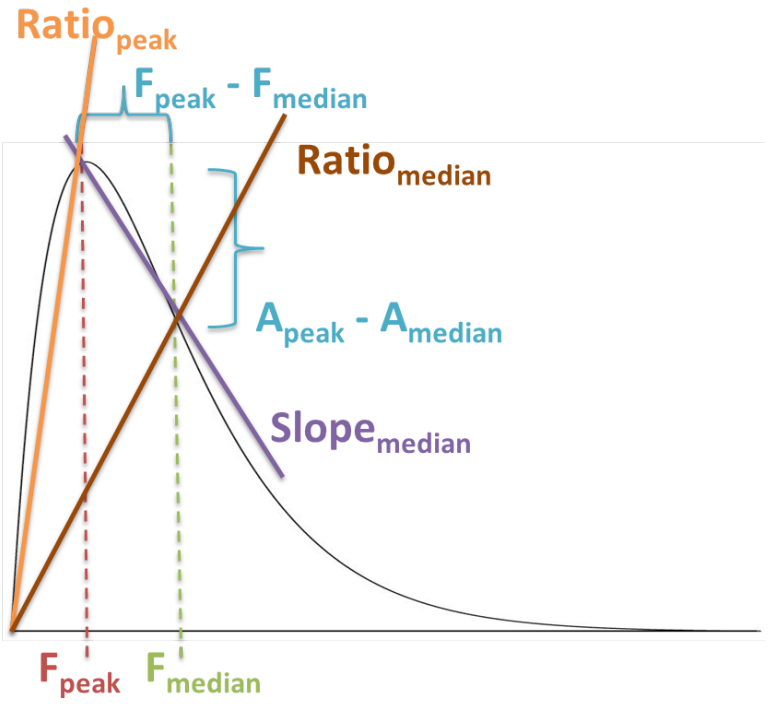

FIGURE 3.10: Definition of spectral F0 features.

3.5.2.1 Spectral analysis of Intonational contours: Statistical analysis

This kind of analysis was performed on bipolar and healthy control subject. In particular, only bipolar subjects who were labelled euthymic in at least one recording were used in this study. Moreover, in this study only the reading task was considered. When the same mood state had been recorded twice for the same speaker, the features extracted from the two audio recordings, were averaged. Thus, for each subject and for each mood state one value for each feature is estimated. Friedman's test was used to check for statistical differences in paired data corresponding to different mood states (in the same patients), while Mann-Whitney U-test was used to investigate such differences for independent samples (i.e. different patients and different mood states). The latter was performed with and without normalization with respect to the same patient's features estimated in the euthymic state. Since the limited number of enrolled bipolar patients did not allow performing a proper approximation of the Friedman F statistics with the χ^2 distribution, the critical values for Friedman's test reported in [19], obtained through simulations, and corresponding to a statistical significance equal to 0.05 were used. Such an approximation was required because the corresponding significance, i.e. α, was not considered reliable.

Inter-session analysis on subjects forming the Healthy Control Subject Database was investigated using a non-parametric Friedman's test for paired data.

3.6 Voice quality

In previous studies the importance of voice quality characterizing phonation in people suffering from depression [213, 221, 310] or when conveying different emotions in verbal communication [147] was highlighted. As opposed to prosody and loudness related features, voice quality was seen as playing a role in communicating the valence of an emotion rather than its activation [311]. Instead, in [147] it was asserted that differences in voice quality can only report different timbres in an otherwise euthymic utterance.

Here, the Long-Term Average Spectrum (LTAS) is estimated to study voice quality from audio signals acquired from bipolar patients and healthy control subjects when reading a neutral text [5]. Moreover, a comparison between the results of the LTAS with and without a correction accounting for differences in speech fundamental frequency (F0), previously tested on synthetic vowels, is introduced [5]. In addition, this method is also applied to the German Emotional Database to test possible statistically significant differences among the four emotions under study.

3.6.1 Long-Term Average Spectrum of Speech

The first step consists in the localization of voiced intervals by means of the method based on the autocorrelation function and on signal energy (Figure 3.4), while in the second step, LTAS of voiced intervals is calculated by means of the Fast Fourier Transform (fft).

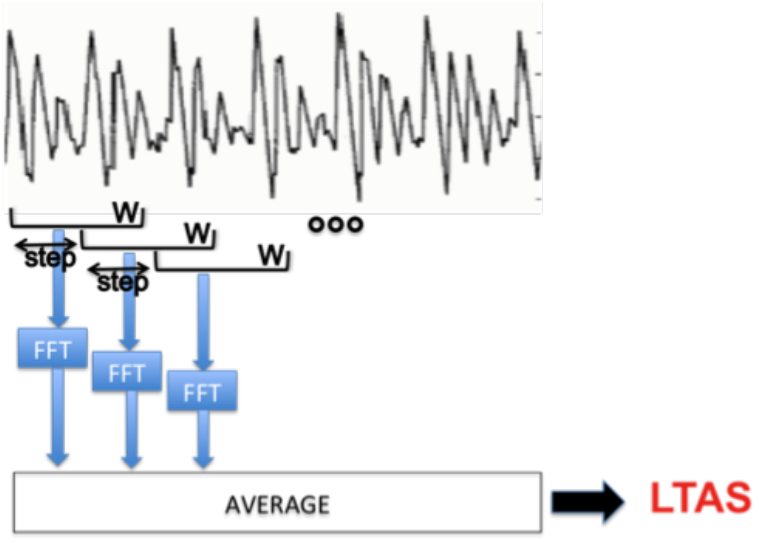

FIGURE 3.11: Scheme of Long-Term Average Spectrum.

All the audio signals are normalized to the range $[-1; 1]$ to compensate for possible intensity changes across recordings. After having localized voiced intervals, the normalized audio signal is filtered by means of a pre-emphasis filter (α=0.99) to reduce ultra-low

frequency spectral components. Then, in a second step, to estimate the LTAS, an fft algorithm is applied within a window sliding within each voiced interval. A Hamming window function of 5 ms long (bin width=200 Hz) and a time hop of 3 ms are used. Finally, the LTAS is estimated by averaging over the whole utterance the amplitude spectra obtained for each voiced frame according to (3.16):

$$L[k] = 1/N \sum_{i=1}^{N} S_i[k] \tag{3.16}$$

where N is the total number of voiced frames, S_i is the amplitude spectrum of the ith frame, and k is the index of the frequency component (see Figure 3.11). Since high-frequency components are characterized by lower amplitudes than low-frequency components, a boosting of the former is performed by calculating the logarithm of the spectral amplitude plus one (3.17).

$$L_{boosted} = log(1 + L) \tag{3.17}$$

3.6.2 F0-corrected LTAS: a proposed method

Since LTAS reflects the contribution of the global source and the vocal tract for the voice quality, a further LTAS estimating algorithm is used to perform a F0 correction of the LTAS. In fact, similarly to [312] windows selecting the current F0 periods are used. Here, the Dypsa algorithm [313] is used to estimate glottal closure instants and thus glottal cycles lengths in each voiced segment. Hence, the frame length is set to the 90% of current F0 period, and the frame start is anticipated, with respect to the glottal closure instant, of the 5% of current F0 period. A longer time-window analysis would provide a spectrum in which the frequencies corresponding to the formants depend also on the glottal impulse [314]. Finally, the obtained spectra are binned in a new frequency axis (bin width=150Hz) before averaging and normalizing to obtain power spectral density in each bin [312]. Such an operation is necessary to properly combine spectra which have a slightly different frequency resolution, given the different windows lengths.

3.6.3 Voice quality study: Method Testing and Statistical Analysis

In this study different kinds of data were analysed: both synthetic and real samples were analysed.

To study the effect of F0 on LTAS, an analysis on synthetic voice data was performed. An autoregressive moving average exogenous ($ARMAX$) model [300] is used to synthesize voice samples at two different F0 mean values and by varying the applied jitter. The model parameters were estimated from a male [a] vowel, using model orders for the AR, MA and X parts equal to 16, 4 and 2 respectively. A comparison between the LTAS profiles estimated from vowels at different F0 is performed.

As regards real voice samples, the results using LTAS estimation with and without F0 correction are discussed in view of the meanF0 distribution across audio signals. MeanF0 is estimated via Camacho's SWIPE'-based algorithm [16]. For this purpose, voiced segments are detected by means of the method based on the autocorrelation function and on signal energy (Figure 3.4), coherently with the one used to estimate LTAS from every voiced segments.

An average LTAS and meanF0 were estimated for all the daily pairs (when two recordings were acquired at a same day) of recordings.

Regarding the German Emotional Database, a non-parametric Kruskal-Wallis test was used to detect possible statistically significant differences in frequency contents related to emotional states in a grouped-subjects analysis. In this case, the Benjamini & Hochberg [315] procedure for controlling the false discovery rate was used to adjust the estimated p-values.

Regarding the Bipolar Database, a non-parametric statistical test was used to investigate differences among the frequency content and meanF0 in different mood states. Friedman's test was used to perform pairwise comparison between euthymia and depression and euthymia and hypomania. Statistical analysis were performed in the band 0-8000 Hz on each frequency bin independently.

Anyway the limited number of enrolled bipolar patients did not allow performing a reliable approximation of the Friedman F statistics with the χ^2 distribution, and of the corresponding significance, i.e. α. To overcome this limitation, the critical values for Friedman's test reported in [19] were used. These critical values were, obtained through simulations, and corresponded to a statistical significance equal to 0.05.

Chapter 4

Results[1]

4.1 Voice activity detection

4.1.1 Proposed VAD method: Parameters settings

Before comparing VAD methods it was important to find the best parameters settings. Several parameter values were investigated to search for the best configuration set. EGG provided by the CMU Arcitic Database [293] were useful, since they enable detecting the instants of maximal acoustic excitation. In Figures 4.1 and 4.2 the obtained specificity and sensitivity values are reported for each parameter set. w is the window length and s is the time hop in ms.

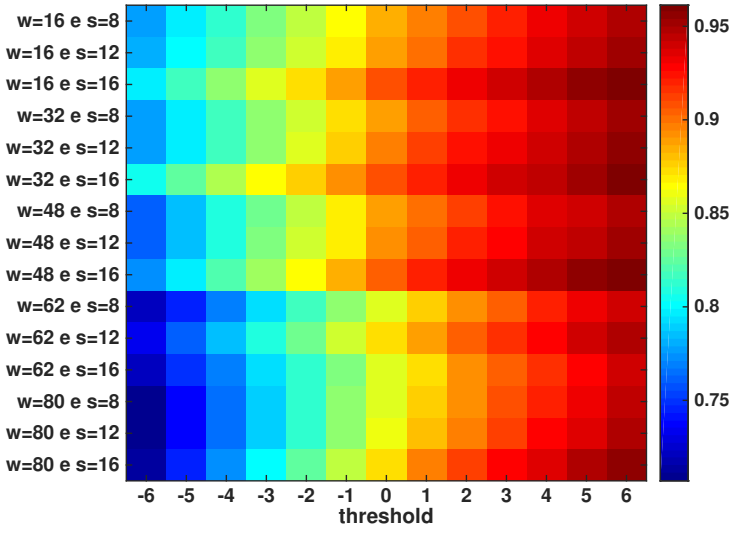

FIGURE 4.1: Specificity values obtained for each configuration set.

[1]Part of this Chapter has been already published in [1–5].

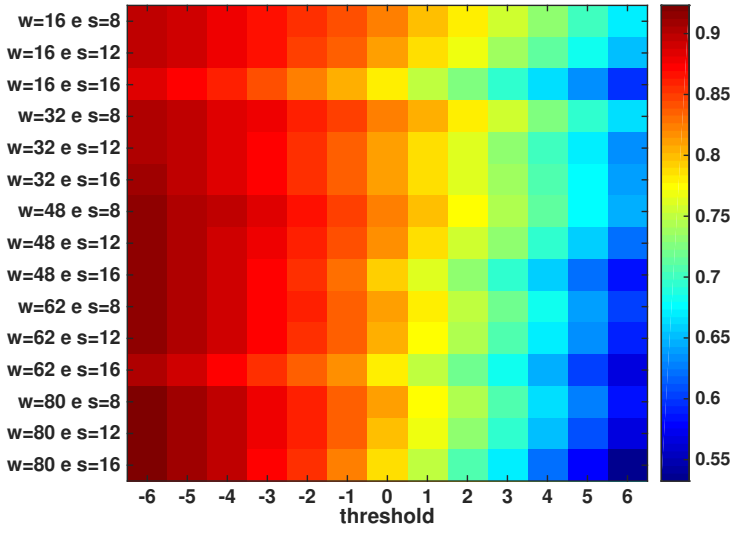

FIGURE 4.2: Sensitivity values obtained for each configuration set.

The best parameter set for the CMU Arctic database, according to the specificity and sensibility criteria, was found to be 16 ms for time hop, 32 ms for the window length and 0 dB with respect to the median intensity level for the threshold.

4.1.2 VAD methods comparison: Test on synthetic data

The time instants detected by the two VAD algorithms were compared with those set during the synthesis process where the target F0 values and vowel lengths were varied. Regarding the benchmark VAD method, the one including autocorrelation functions and signal energy, the results are reported in Figures 4.3 and 4.4. In Figure 4.3 the median absolute value of the error in detecting the starting time (Δt_s) and ending time (Δt_s), and the difference between real and measured vowel lengths (ΔL) are reported when the target F0 value is varied. The same statistics are reported in Figure 4.4 varying the vowel length.

The same statistics concerning the proposed VAD algorithm, i.e. the algorithm involving signal intensity and zero crossing rate, are reported in Figures 4.5 and 4.6.

Both algorithms show similar trends: a lower absolute median error in the quicker transitions, i.e. at the beginning of the first vowels and at the ending of the second ones. While a greater absolute median error occur in the slower transitions. Moreover, such errors seem to increase with the vowel length and thus with the slowness of the transitions (see Figures 4.4 and 4.6).

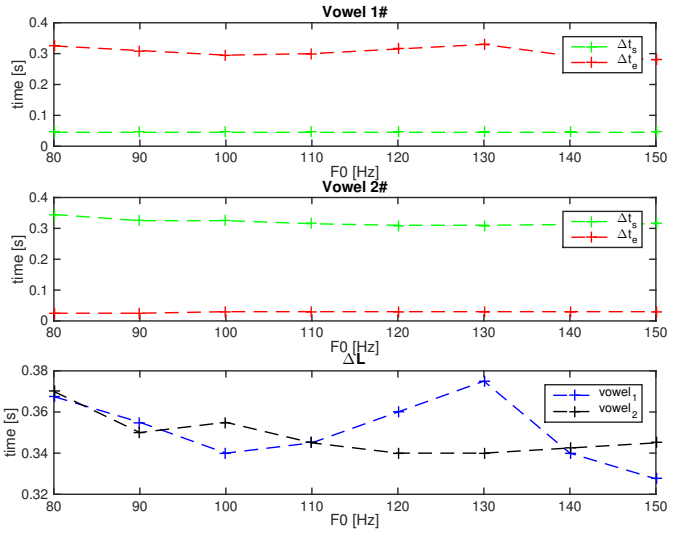

FIGURE 4.3: VAD - Benchmark method: Trends of Δt_s, Δt_e, and ΔL when varying F0.

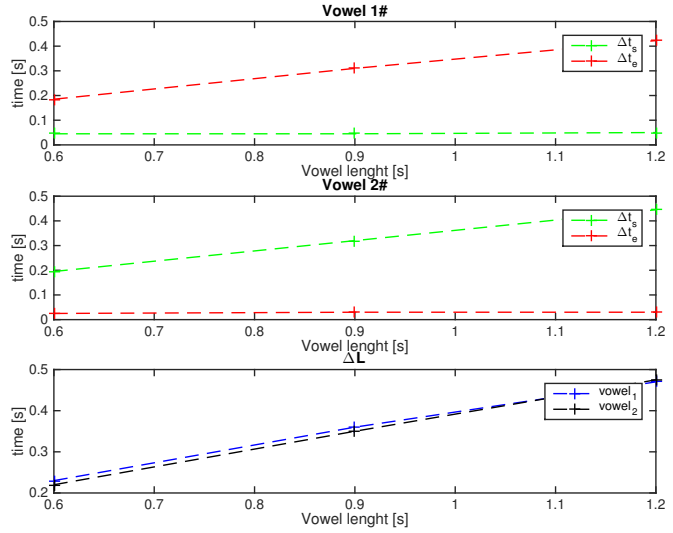

FIGURE 4.4: VAD - Benchmark method: Trends of Δt_s, Δt_e, and ΔL when varying Vowel length.

TABLE 4.1: Mean absolute error between time intervals [s].

	Vowel 1		**Vowel 2**	
	Δt_s	Δt_e	Δt_s	Δt_e
Benchmark Method	0.045	0.310	0.320	0.030
Proposed Method	0.020	0.240	0.215	0.025

The comparison between the two methods enables to conclude that the proposed method outperforms the benchmark method since lower absolute mean differences between real and detected transition time intervals are obtained (Table 4.1). In Figures 4.7 and 4.8

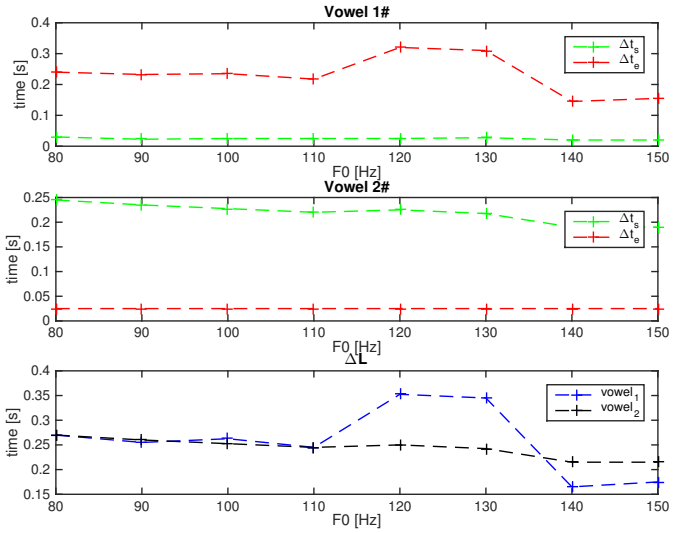

FIGURE 4.5: VAD - Proposed method: Trends of Δt_s, Δt_e, and ΔL when varying F0.

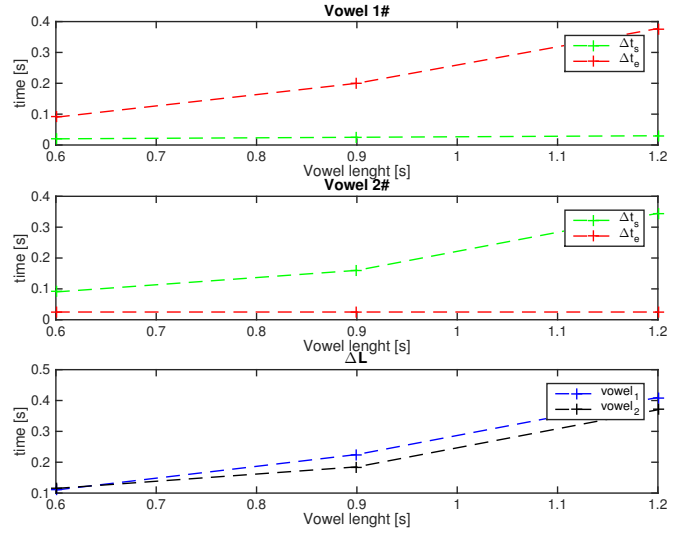

FIGURE 4.6: VAD - Proposed method: Trends of Δt_s, Δt_e, and ΔL when varying Vowel length.

the relative histograms are displayed.

4.2 F0 estimation algorithm: Test on synthetic data

To evaluate the reliability of the F0 estimation algorithm, the percentage deviations from the expected F0 values were taken into account. Table 4.2 reports the median percentage deviation from F0 target values. The median percentage deviations from the expected F0 values are low. A median percentage error, for all the five target times,

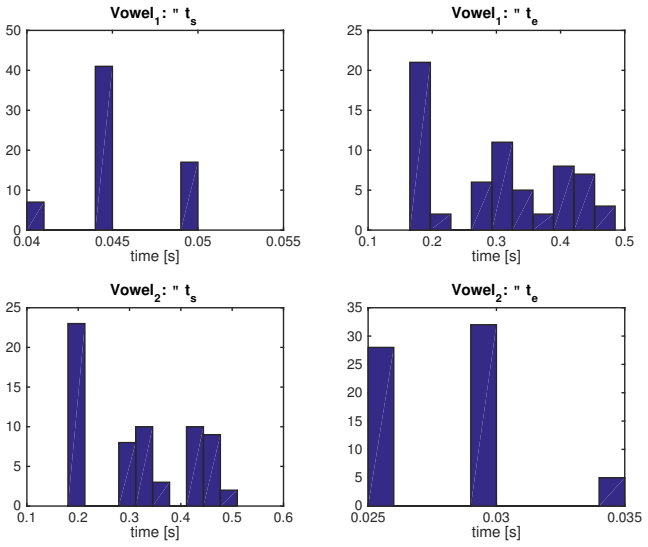

FIGURE 4.7: VAD - Benchmark method: Histogram of median absolute errors.

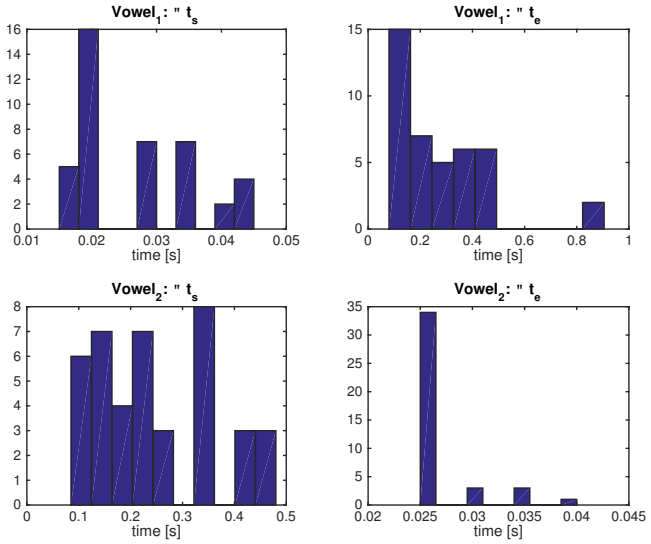

FIGURE 4.8: VAD - Proposed method: Histogram of median absolute errors.

TABLE 4.2: Percentage deviations from the expected F0 values [%]. The median values are reported according to F0 trajectory and globally.

F0 trajectories	$F0_1$	$F0_2$	$F0_3$	$F0_4$	$F0_5$
timing [s]	0	1.2	1.5	2.4	3.0
rising	1.14E-01	1.58E-01	4.21E-01	2.90E-01	9.53E-01
peak	1.14E-01	1.57E-01	4.07E-01	1.82E-01	7.38E-01
falling	1.59E-01	2.24E-01	4.67E-01	1.13E-01	7.03E-01
valley	1.59E-01	2.25E-01	4.02E-01	3.10E-01	6.27E-01
flat	1.36E-01	2.34E-01	4.37E-01	2.02E-01	5.82E-01
total	1.14E-01	2.02E-01	4.27E-01	2.63E-01	7.03E-01

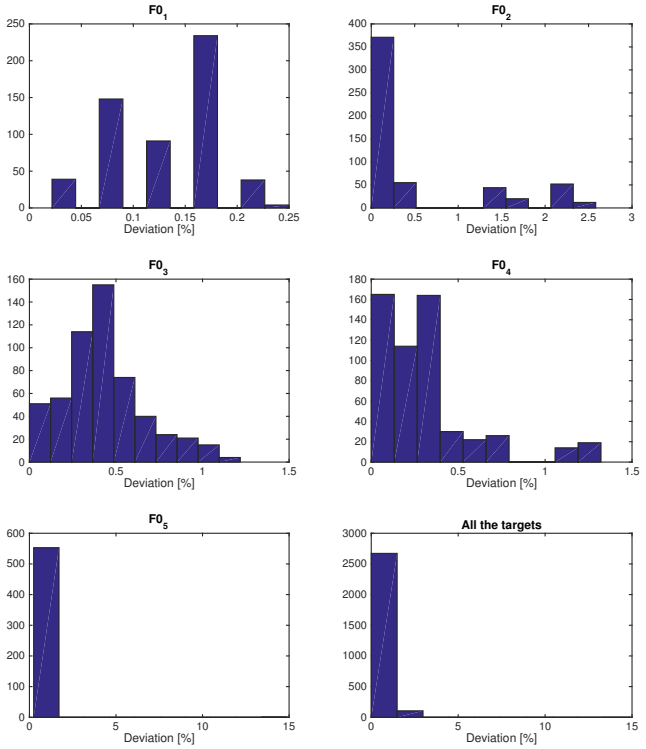

FIGURE 4.9: Histograms of the percentage deviations from the expected F0 values. The indices refer to the target times.

equal to $2.85E-1$ % was reported. The corresponding histograms are shown in Figure 4.9.

4.3 Bipolar dataset: scoring

The subjects were in a different mood state for all the daily recording sessions (see Tables 4.3 and 4.4). In the first recording session, subjects A, B, C and G were scored as hypomanic while subjects D, E, F, H, I and L were scored as depressed, and subject G was scored as mixed. All subjects but two were scored as euthymic in the second recording session. Subject E was scored as hypomanic in the second acquisition day, and subject M as depressed. Subject B was scored as depressed and subject M as euthymic in the third session.

TABLE 4.3: Patients suffering from bipolar disease enrolled in Strasbourg.

subj.	gender	age	label day 1	label day 2	label day 3
A	M	40	Hypomania	Euthymia	
B	M	53	Hypomania	Euthymia	Depression
C	M	40	Hypomania	Euthymia	
D	M	28	Depression	Euthymia	
E	F	34	Depression	Hypomania	
F	M	54	Depression	Euthymia	
G	F	37	Hypomania	Euthymia	

TABLE 4.4: Patients suffering from bipolar disease enrolled in Pisa.

subj.	gender	age	label day 1	label day 2	label day 3
H	F	32	Depression	Euthymia	
I	M	52	Depression	Euthymia	
L	F	36	Depression	Euthymia	
M	F	34	Mixed	Depression	Euthymia

4.4 Vocal features

4.4.1 F0, F0 standard deviation, frame-to-frame jitter

4.4.1.1 Vocal features: Tests on CMU Arctic Database

The F0 estimates obtained from the EGG signal were compared with those obtained from the audio files. The speech corpus considered for the test contains more than 8000 vowels. A linear regression model was adopted for meanF0, while a logarithmic model was considered when evaluating the F0 standard deviation (stdF0), the cycle-to-cycle jitter and the frame-to-frame jitter (LpJ). The voiced segment average F0 (meanF0) is always very well correlated with F0 estimated from the EGG files. The slope in the log-log regression model relates the percentage changes in the dependent variable to those in the independent variable. Thus, a slope smaller than one identifies a smaller increase in the estimated features with respect to the actual data measured from EGG data. Both stdF0 and LpJ cannot be reliably estimated when echo is applied and they are more sensitive to noise than average F0. Cycle-to-cycle jitter is less correlated with the true values than LpJ.

The correlation coefficient and the slope of the regression models are reported in Table 4.5. In particular, the correlation coefficient between the two values of F0 is 0.99 and the fitted linear regression model has a slope equal to 1.02 and an intercept equal to $-2.9\ Hz$. Figure 4.10 shows the corresponding graph.

TABLE 4.5: Slope (α) and correlation coefficient (ρ): regression model relating the features from audio and EGG files.

	F0		**SD**		**cycle-to-cycle Jitter**		**LpJ**	
	Linear model		**Log-Log. model**		**Log-Log. model**		**Log-Log. model**	
	α	ρ	α	ρ	α	ρ	α	ρ
No Noise	1.02	0.99	1.00	0.95	0.62	0.79	0.98	0.94
Street Noise	1.02	0.99	1.00	0.93	0.63	0.78	0.98	0.90
Train Noise	1.03	0.99	1.00	0.92	0.62	0.77	0.97	0.90
Gaussian Noise 15dB	1.02	0.99	1.00	0.94	0.63	0.80	0.99	0.93
Gaussian Noise 10dB	1.02	0.99	1.00	0.94	0.60	0.78	0.98	0.93
Gaussian Noise 5dB	1.03	0.99	0.98	0.90	0.55	0.75	0.88	0.87
Echo 20ms	1.01	0.99	0.66	0.71	0.36	0.40	0.62	0.63
Echo 50ms	1.03	0.97	0.58	0.61	0.37	0.40	0.58	0.57
Echo 100ms	1.05	0.95	0.64	0.66	0.45	0.55	0.62	0.66

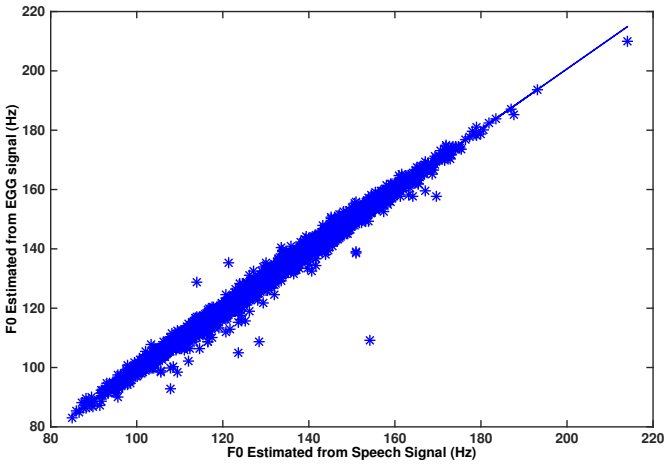

FIGURE 4.10: F0 estimated from audio signals (x-axis) compared with F0 from EGG signal.

4.4.1.2 Vocal features: Emotion Database

The results on the intra-subject analysis are reported in Tables 4.6, 4.7 and 4.8. Statistically significant differences between every couple of emotion are highlighted with different symbols, namely: $*$ (Anger versus Neutral), $+$ (Anger versus Boredom), ¶ (Anger versus Happiness), $\Diamond$ (Neutral versus Boredom), ∂ (Neutral versus Happiness) and $\bullet$ (Boredom versus Happiness). The greatest number of statistically significant differences concerns the comparisons between the emotion related to higher and lower arousal states, i.e. between boredom or neutral and anger or happiness. In some subjects it is possible to detect some statistically significant differences between the features extracted from audio recorded while the actors were playing higher arousal emotions, i.e happiness vs. anger, or lower arousal emotions, i.e. boredom vs. neutral.

Such behaviour is observable on the three investigated features: average F0 (meanF0), standard deviation of F0 (stdF0), and frame-to-frame jitter (LpJ). Anyway, meanF0

TABLE 4.6: Mean and SD of F0 estimated from voiced segments (Hz).

meanF0				
Subj..	**Anger**	**Neutral**	**Boredom**	**Happiness**
1	215.26 ± 35.22 ∗+	125.32 ± 12.63 ∗◊∂	109.49 ± 12.02 +◊•	221.08 ± 33.45 ∂•
2	303.45 ± 41.68∗+¶	207.72 ± 23.32 ∗◊∂	188.58 ± 31.82 +◊	262.79 ± 48.06 ¶∂•
3	279.07 ± 39.44 ∗+¶	163.96 ± 14.10 ∗∂	167.99 ± 21.67 +•	339.33 ± 50.70 ¶∂•
4	193.22 ± 27.63 ∗+	109.99 ± 10.42 ∗∂	105.98 ± 9.97 +•	213.92 ± 23.63 ∂•
5	216.40 ± 40.59 ∗+¶	113.52 ± 8.18 ∗∂	112.21 ± 11.40 +•	188.59 ± 28.00 ¶∂•
6	212.79 ± 24.50 ∗+¶	138.50 ± 11.43 ∗◊	147.56 ± 21.70 +◊	144.29 ± 12.25 ¶
7	319.37 ± 37.06 ∗+	198.84 ± 16.98 ∗◊∂	171.56 ± 20.53 +◊•	304.69 ± 46.96 ∂•
8	294.90 ± 39.34 ∗+¶	167.09 ± 12.90 ∗◊∂	176.45 ± 33.73 +◊•	283.41 ± 37.76 ¶∂•
9	224.44 ± 27.06 ∗+	107.22 ± 11.00 ∗∂	101.54 ± 12.58 +•	228.58 ± 48.28 ∂•
10	309.26 ± 46.68 ∗+¶	202.98 ± 25.32 ∗∂	201.20 ± 36.12 +•	335.27 ± 49.47 ¶∂•

TABLE 4.7: Mean and SD of F0 standard deviation estimated from voiced segments (Hz).

stdF0				
Subj.	**Anger**	**Neutral**	**Boredom**	**Happiness**
1	11.69 ± 6.60 ∗+	4.46 ± 2.92 ∗∂	6.25 ± 4.31 +•	15.11 ± 10.01 ∂•
2	22.95 ± 11.73 ∗+¶	10.04 ± 4.78 ∗◊∂	5.43 ± 3.58 +◊•	15.86 ± 10.12 ¶∂•
3	13.28 ± 8.23 ∗+	6.49 ± 3.82 ∗◊∂	3.83 ± 2.25 +◊•	14.04 ± 10.54 ∂•
4	13.84 ± 7.71 ∗+	3.23 ± 2.49 ∗◊∂	2.22 ± 1.49 +◊•	10.70 ± 7.00 ∂•
5	15.44 ± 9.10 ∗+	3.16 ± 2.29 ∗∂	2.91 ± 1.37 +•	13.11 ± 7.92 ∂•
6	10.91 ± 6.88 ∗+	5.26 ± 3.11 ∗◊	3.79 ± 2.59 +◊•	8.71 ± 5.81 •
7	14.79 ± 0.69 ∗+	7.17 ± 4.45 ∗∂	5.46 ± 3.73 +•	17.81 ± 10.08 ∂•
8	16.17 ± 8.07 ∗+	5.30 ± 2.61 ∗∂	6.35 ± 4.06 +•	19.39 ± 10.25 ∂•
9	10.83 ± 6.86 ∗+	2.51 ± 1.77 ∗∂	2.61± 1.79 +•	12.35 ± 5.51 ∂•
10	16.22 ± 8.86 ∗+	5.20 ± 3.05 ∗∂	6.77 ± 3.54 +•	17.29 ± 8.53 ∂•

TABLE 4.8: Mean and SD of frame-to-frame jitter estimated from voiced segments (%).

LpJ				
Subj.	**Anger**	**Neutral**	**Boredom**	**Happiness**
1	1.19 ± 0.53	1.03 ± 0.57	1.39 ± 0.98	1.40 ± 0.64
2	1.10 ± 0.38 +	1.03 ± 0.32 ◊	0.57 ± 0.26 +◊•	1.03 ± 0.40 •
3	0.81 ± 0.33 ∗	0.96 ± 0.51 ∗◊∂	0.67 ± 0.35 ◊	0.65 ± 0.40 ∂
4	1.46 ± 0.52 ∗+	1.24 ±0.66 ∗◊	0.79 ± 0.38 +◊•	1.33 ± 0.70 •
5	1.32 ± 0.56 ∗+	1.23 ± 0.70 ∗∂	0.82 ± 0.39 +•	1.61 ± 0.75 ∂•
6	1.03 ± 0.49 +¶	1.08 ± 0.59 ◊	0.68 ± 0.38 +◊•	1.71 ± 1.04 ¶•
7	0.88 ± 0.42 +	0.80 ± 0.35 ◊	0.63 ± 0.27 +◊•	0.84 ± 0.37 •
8	0.96 ± 0.40 +¶	0.88 ± 0.32 ◊∂	0.73 ± 0.33 +◊•	1.12 ± 0.52 ¶∂•
9	0.96 ± 0.35 ¶	0.99 ± 0.61	0.87 ± 0.40 •	1.23 ± 0.63 ¶•
10	0.72 ± 0.27 +	0.70 ± 0.32	0.59 ± 0.23 +•	0.74 ± 0.27 •

seems to outperform the other two features, since in six out of ten subjects it displays statistically significant differences between the two high arousal emotions, and in five out of ten between the two low arousal emotions.

The results regarding the groupwise analysis are reported in Figures 4.11, 4.12, and

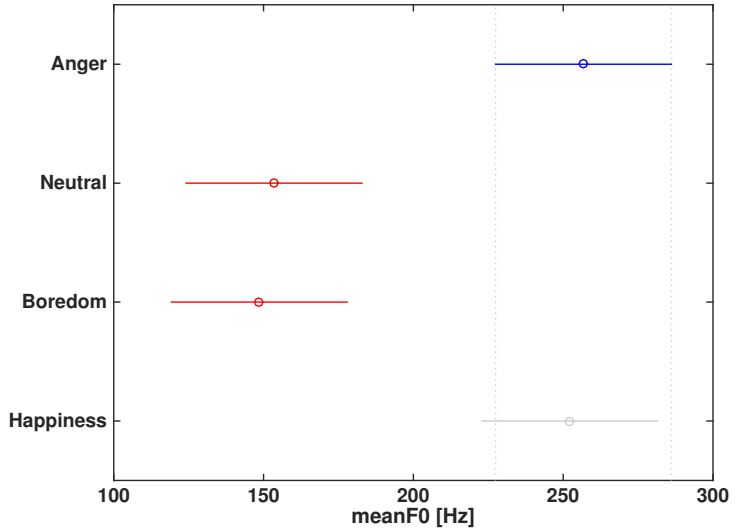

FIGURE 4.11: Results at group level of emotional speech data. Graphs of one-way ANOVA test of meanF0.

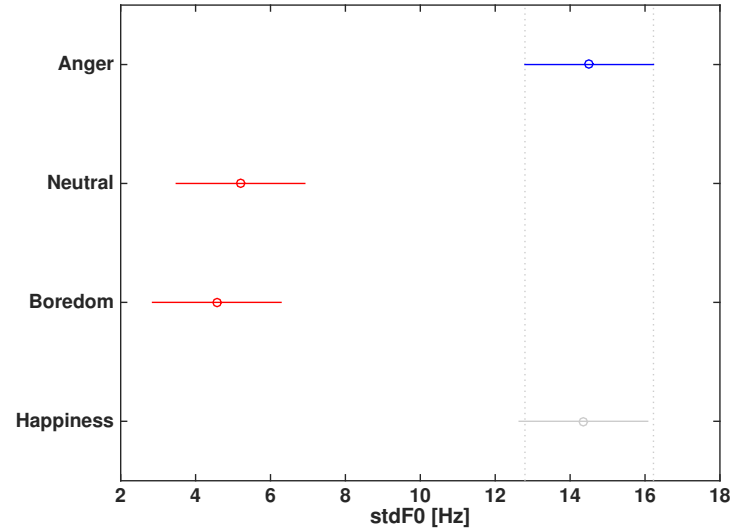

FIGURE 4.12: Results at group level of emotional speech data. Graphs of one-way ANOVA test of stdF0.

TABLE 4.9: Group results of emotional speech: average values and p-values of the one-way ANOVA tests.

		Mean			
feature	**p-value**	**anger**	**neutral**	**boredom**	**happiness**
meanF0	1,64e-06	256.78	153.48	148.54	252.23
stdF0	3,52e-11	14.50	5.20	4.56	14.35
Jitter	1,38E-02	1.03	0.99	0.77	1.16

4.13 and in Table 4.9. The analyses confirm for meanF0 and stdF0 the statistically significant differences between levels of arousal. In fact, such features report to be

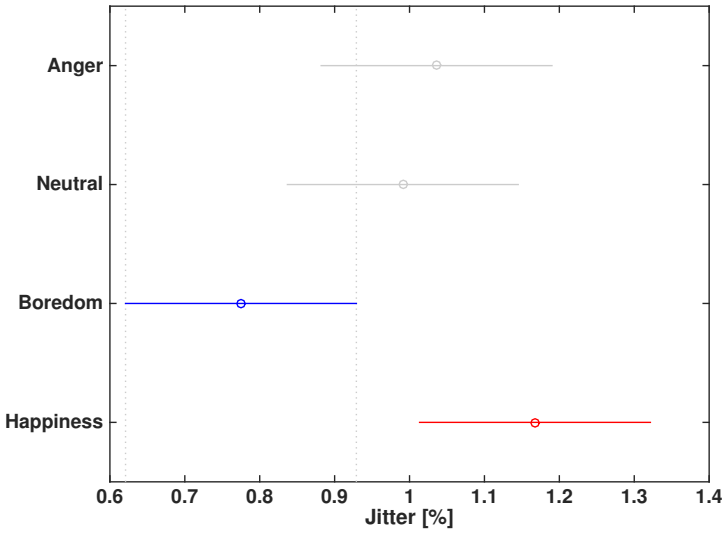

FIGURE 4.13: Results at group level of emotional speech data. Graphs of one-way ANOVA test of Jitter.

statistically significant different in high arousal emotion recordings (happiness and anger) with respect to the low arousal emotion recordings (boredom and neutral). LpJ shows statistically significant differences between boredom and happiness only.

4.4.1.3 Vocal features: Bipolar Database and Healthy Control Subject Database

In this section, preliminary results on patients are described. Only the recordings of the first two acquisition days are taken into account here. All the subjects were in a different mood state in the second recording session with respect to the first one. F0 values have been normally distributed, while frame-to-frame jitter and within voiced segment F0 standard deviations, have not been normally distributed. Results are reported in Tables 4.10 to 4.12. Statistically significant differences between recordings related to the same task are marked with the same symbol (* or †).

MeanF0 was found to be higher in the hypomanic state with respect to euthymic state in the reading task (Table 4.10). When statistically significantly different, the same trend was found in TAT task, except for subject G. Subject G did not show a coherent trend in the two tasks. Moreover, F0 was higher in the euthymic compared to the depressed state for subjects E, H and I, while the opposite was observed in subjects F and L. Differences were observed between tasks (Table 4.10). Subject F, passing from depression to euthymia, did not show a coherent trend between the two tasks, while subject M, passing from mixed to depressed state, showed a lower meanF0 in the first state with respect to the latest one, in both tasks (Table 4.10).

TABLE 4.10: Mean and standard deviation (SD) of meanF0 estimated from voiced segments (Hz).

Mood States		**meanF0**				
Subj.	**Day 1**	**Day 2**	**Reading 1**	**TAT 1**	**Reading 2**	**TAT 2**
A	Hypomania	Euthymia	**130 ± 7** ∗	130 ± 17	**125 ± 8** ∗	129 ± 21
B	Hypomania	Euthymia	**112 ± 9** ∗	**132 ± 24** ∗	**109 ± 10** ∗	**104 ± 14** ∗
C	Hypomania	Euthymia	**100 ± 8** ∗	103 ± 15	**97 ± 6** ∗	102 ± 13
D	Depression	Euthymia	106 ± 4	106 ± 11	105 ± 4	105 ± 11
E	Depression	Hypomania	**189 ± 9** ∗	179 ± 29	**192 ± 10** ∗	178 ± 36
F	Depression	Euthymia	**107 ± 4** ∗	**117 ± 9** †	**101 ± 7** ∗	**100 ± 11** †
G	Hypomania	Euthymia	**202 ± 18** ∗	**188 ± 39** †	**195 ± 18** ∗	**206 ± 47** †
H	Depression	Euthymia	**189 ± 19** ∗	**184 ± 15** †	**215 ± 30** ∗	**210 ± 33** †
I	Depression	Euthymia	**124 ± 13** ∗	**122 ± 17** †	**141 ± 17** ∗	**132 ± 25** †
L	Depression	Euthymia	**182 ± 83** ∗	157 ± 28	**172 ± 19** ∗	159 ± 27
M	Mixed	Depression	**238 ± 24** ∗	**237 ± 26** †	**243 ± 18** ∗	**243 ± 26** †

TABLE 4.11: Median and median absolute deviation (MAD) of stdF0 estimates (Hz).

Mood States		**stdF0**				
Subj.	**Day 1**	**Day 2**	**Reading 1**	**TAT 1**	**Reading 2**	**TAT 2**
A	Hypomania	Euthymia	2.83 ± 1.80	3.27 ± 1.91	3.04 ± 1.88	3.13 ± 2.08
B	Hypomania	Euthymia	4.60 ± 3.40	**4.88 ± 3.55** ∗	4.81 ± 3.76	**2.58 ± 1.67** ∗
C	Hypomania	Euthymia	**2.97 ± 1.57** ∗	**2.85 ± 1.82** †	**2.50 ± 1.35** ∗	**2.36 ± 1.34** †
D	Depression	Euthymia	1.97 ± 1.10	2.25 ± 1.28	1.93 ± 1.04	2.19 ± 1.17
E	Depression	Hypomania	4.10 ± 1.93	5.81 ± 3.40	4.19 ± 2.22	6.30 ± 3.93
F	Depression	Euthymia	2.32 ± 1.17	**2.87 ± 1.76** †	2.38 ± 1.27	**2.68 ± 1.52** †
G	Hypomania	Euthymia	**6.32 ± 3.55** ∗	**4.53 ± 2.64** †	**5.06 ± 2.74** ∗	**5.99 ± 3.75** †
H	Depression	Euthymia	3.89 ± 1.91	**3.80 ± 2.03** †	3.70 ± 1.91	**7.97 ± 3.43** †
I	Depression	Euthymia	3.37 ± 2.15	2.35 ± 1.65	2.94 ± 1.60	2.58 ± 1.83
L	Depression	Euthymia	4.21 ± 2.85	1.85 ± 1.15	4.55 ± 3.05	3.08 ± 2.33
M	Mixed	Depression	**4.56 ± 2.47** ∗	3.09 ± 1.66	**3.92 ± 2.03** ∗	3.21 ± 1.86

When statistically significant differences were found (Table 4.11), the stdF0 estimated from the recordings in the hypomanic state were higher than those observed in the euthymic state, without exception. Subject G showed the opposite trend in the TAT task. Incoherent trends were find concerning the depression-hypomania transitions. A higher stdF0 value was observed in mixed state with respect to the depressed one.

TABLE 4.12: Median and median absolute deviation (MAD) of jitter estimated from voiced segments (%).

Mood States		**LpJ**				
Subj.	**Day 1**	**Day 2**	**Reading 1**	**TAT 1**	**Reading 2**	**TAT 2**
A	Hypomania	Euthymia	**0.58 ± 0.30** ∗	0.73 ± 0.38	**0.70 ± 0.36** ∗	0.69 ± 0.38
B	Hypomania	Euthymia	1.13 ± 0.70	**1.04 ± 0.65** ∗	1.12 ± 0.72	**0.80 ± 0.45** ∗
C	Hypomania	Euthymia	**0.94 ± 0.43** ∗	**0.83 ± 0.47** †	**0.86 ± 0.41** ∗	**0.69 ± 0.36** †
D	Depression	Euthymia	0.62 ± 0.27	0.67 ± 0.35	0.64 ± 0.27	0.70 ± 0.34
E	Depression	Hypomania	0.50 ± 0.20	0.67 ± 0.34	0.49 ± 0.22	0.71 ± 0.43
F	Depression	Euthymia	**0.66 ± 0.30** ∗	**0.64 ± 0.34** †	**0.77 ± 0.35** ∗	**0.82 ± 0.40** †
G	Hypomania	Euthymia	**0.62 ± 0.29** ∗	**0.49 ± 0.23** †	**0.56 ± 0.25** ∗	**0.59 ± 0.32** †
H	Depression	Euthymia	**0.55 ± 0.23** ∗	0.53 ± 0.31	**0.40 ± 0.16** ∗	0.83 ± 0.22
I	Depression	Euthymia	0.72 ± 0.37	0.51 ± 0.30	0.62 ± 0.26	0.54 ± 0.37
L	Depression	Euthymia	0.54 ± 0.30	0.32 ± 0.17	0.64 ± 0.37	0.54 ± 0.35
M	Mixed	Depression	0.44 ± 0.18	0.37 ± 0.19	0.42 ± 0.18	0.33 ± 0.15

In two out of three patients, showing statistically significant differences, LpJ was found to be higher in the hypomanic state compared to the euthymic one (Table 4.12). The direction of change was not always coherent. Subject F showed, coherently in both tasks, an LpJ lower in the depressed state with respect to the euthymic state (Table 4.12). The opposite trend was observed for subject H in the reading task.

It was possible to perform pairwise comparisons exploiting 4 hypomania-euthymia and 5 depression-euthymia transitions (Table 4.13). Statistically significant differences were observed in meanF0 between hypomania and euthymia in the neutral reading task. A decrease of F0 was observed. With regard to LpJ a coherent increase was observed for the depression-euthymia transition.

TABLE 4.13: Results regarding paired inter-state analysis on bipolar data. In bold the statistically significant differences are highlighted. The significant differences are detected according to the critical values for Friedman's F_r reported in [19].

P-values	**HE - Friedman's test**		**DE - Friedman's test**	
	Reading	**TAT**	**Reading**	**TAT**
meanF0	**<0.05**	>0.05	>0.05	>0.05
LpJ	>0.05	>0.05	>0.05	**>0.05**
stdF0	>0.05	>0.05	>0.05	>0.05

The analysis on the Healthy Control Subjects Database, performed by means of a non-parametric Friedman's test for paired data, revealed no statistically significant differences between the two acquisition days (Table 4.14).

TABLE 4.14: Results regarding paired inter-state analysis on Healthy Control Subjects Database. In bold the statistically significant p-values are highlighted.

P-values	**Day1 vs. Day2**	
	Reading	**TAT**
meanF0	6.37E-01	2.06E-01
LpJ	0.59E-01	2.06E-01
stdF0	6.37E-01	2.06E-01

4.5 Prosodic features

4.5.1 Taylor's Extended Intonational Model

All features estimated for single subjects are not normally distributed according to a Lilliefors test. The Mann-Whitney U-test is therefore used for intra-subject analysis. For the group analysis a Kruskal-Wallis test is used to examine possible differences among conditions in ampl*, dur* and tilt*. A one-way ANOVA is used with PosSlope, AbsNegSlope, SumDer and GlobalSlope since at group level they are normally distributed.

4.5.1.1 Taylor's Extended Intonational Model: Emotion Database

Both intra subject analysis (data not shown) and group analysis show statistically significant differences among emotions characterized by high arousal with respect to low arousal (happiness and anger vs. boredom and neutral). With regard to intra-subject analysis, in some subjects statistically significant differences are observed between neutral and boredom, while no differences are observed between anger and happiness.

The group analysis shows that ampl* and tilt* enable distinguishing anger and happiness from boredom and neutral. Dur* enables distinguishing boredom from happiness and anger.

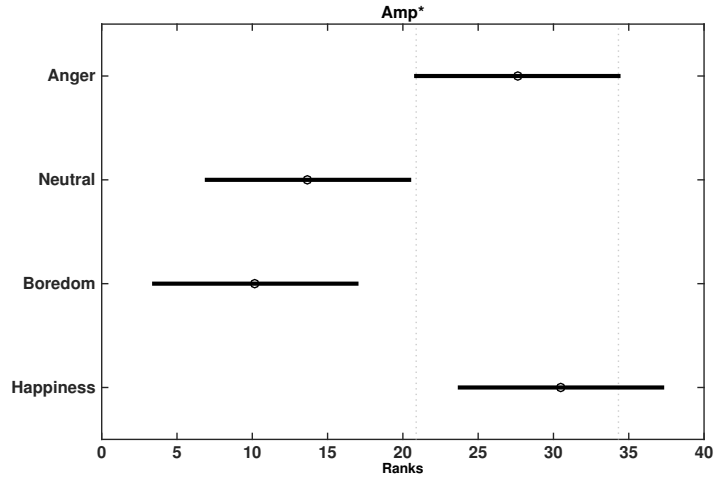

FIGURE 4.14: Results at group level for emotional speech. Graphs of Kruskal-Wallis test of Ampl*.

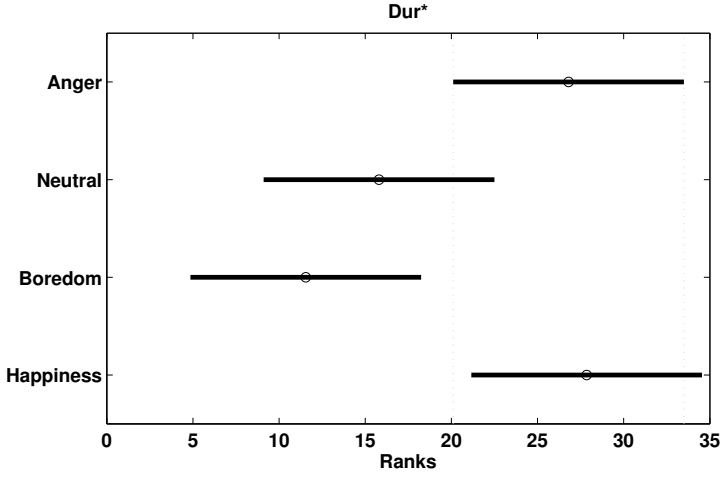

FIGURE 4.15: Results at group level for emotional speech. Graphs of Kruskal-Wallis test of Dur*.

Figures 4.14, 4.15 and 4.16 show the results of the Kruskal-Wallis test. For each group, mean ranks are marked by a circle and an interval equal to the rank MAD. If two intervals

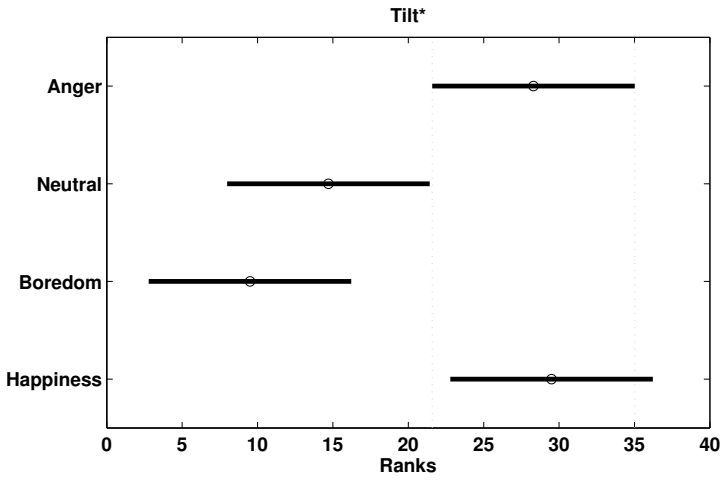

FIGURE 4.16: Results at group level for emotional speech. Graphs of Kruskal-Wallis test of Tilt*.

are disjoint, the groups are significantly different. If the intervals overlap, the groups do not differ significantly. The graphs reporting the one-way ANOVA applied to PosSlope, AbsNegSlope, SumDer and GlobalSlope are also interpretable. PosSlope (Figure 4.17), AbsNegSlope (Figure 4.18), SumDer (Figure 4.19) and GlobalSlope (Figure 4.20) enable distinguishing anger and happiness from boredom and neutral.

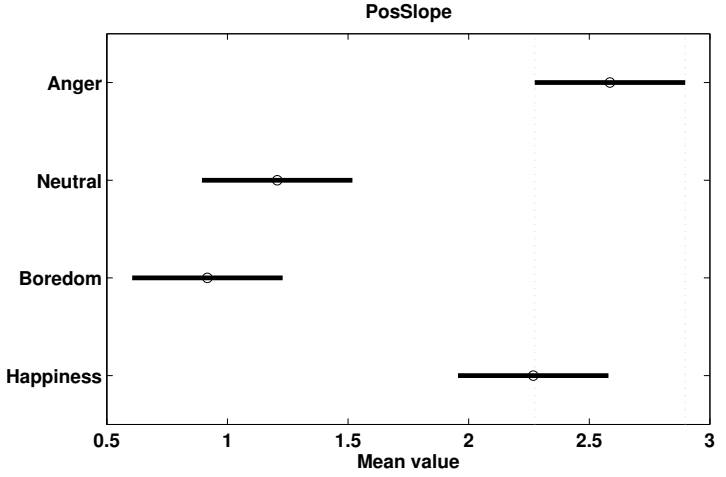

FIGURE 4.17: Results at group level for emotional speech data. Graphs of one-way ANOVA test of PosSlope.

In Table 4.15, the p-values of the tests and the median or mean of each group are reported (see also Figure 4.21). Emotional speech characterized by a low arousal is described by features with a lower median value than emotional speech characterized by higher arousal.

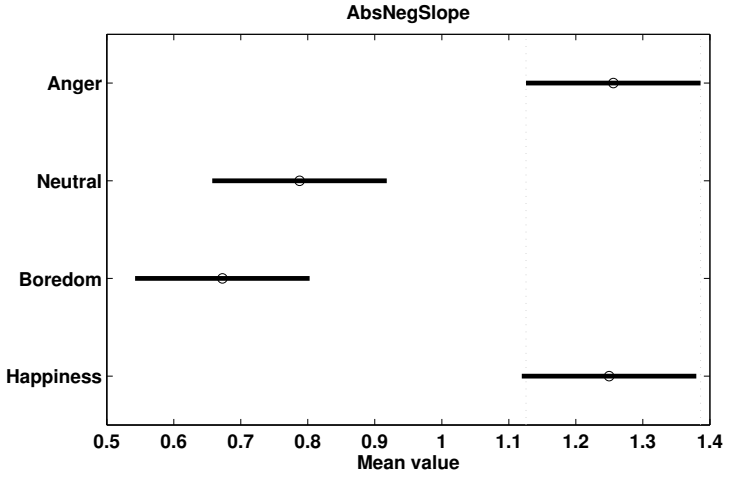

FIGURE 4.18: Results at group level for emotional speech data. Graphs of one-way ANOVA test of AbsNegSlope.

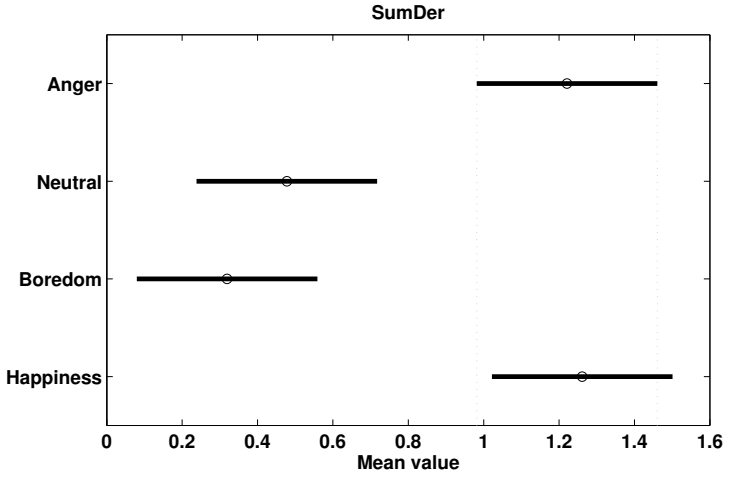

FIGURE 4.19: Results at group level for emotional speech data. Graphs of one-way ANOVA test of SumDer.

TABLE 4.15: Inter-subject analysis results of emotional speeches. Median values and p-values of the Kruskal-Wallis tests are shown.

		median			
feature	**p-value**	**anger**	**neutral**	**boredom**	**happiness**
Amplitude*	6.07E-05	0.47	-0.01	-0.16	0.56
Duration*	2.39E-03	-0.01	-0.27	-0.39	0.05
Tilt*	7.55E-05	0.18	-0.15	-0.25	0.25
PosSlope	8.96E-09	2.58	1.2	0.91	2.27
AbsNehSlope	8.72E-08	1.25	0.78	0.67	1.24
SumDer	1.46E-06	1.22	0.47	0.31	1.26
GlobalSlope	2.40E-08	1.17	0.44	0.16	1.2

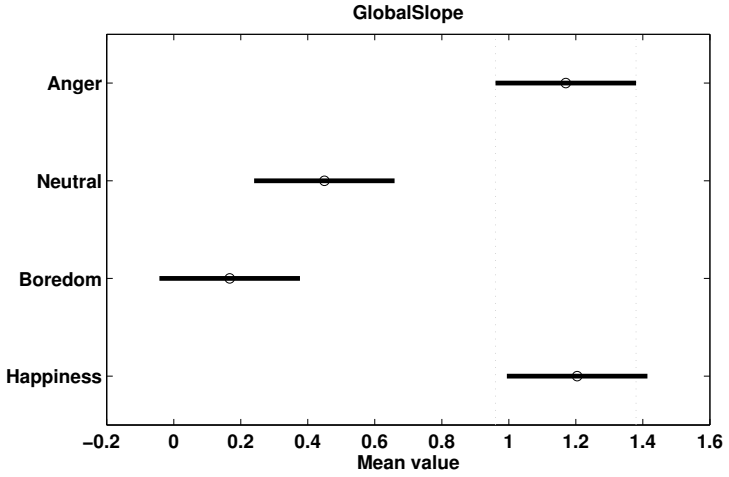

FIGURE 4.20: Results at group level for emotional speech data. Graphs of one-way ANOVA test of GlobalSlope.

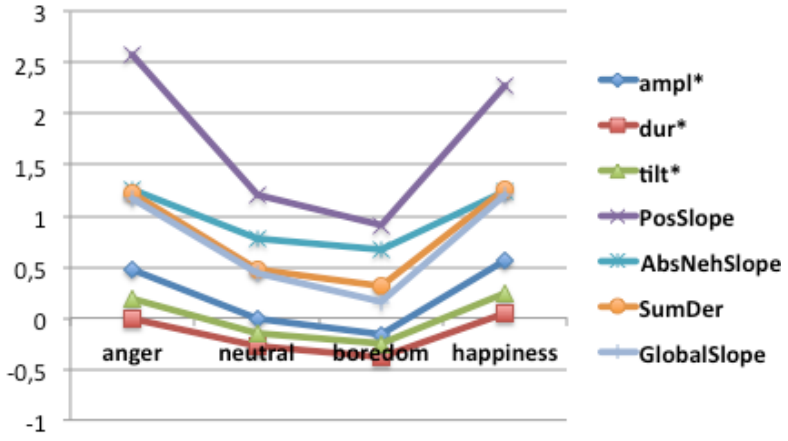

FIGURE 4.21: Median or mean of each group.

4.5.1.2 Taylor's Extended Intonational Model: Bipolar Database

For that database, an intra-subject analysis was performed. Only the features related to the same task were compared. For the reading task, seven patients out of eleven show statistically significant differences for the ampl* feature (Table 4.16). In three patients out of four the median value was found to decrease from hypomania to euthymia. The ampl* value was also observed to be higher in hypomania than in depression. In two out of five patients, an increase in ampl* median value is observed passing from depression to euthymia; while in the other three subjects no statistically significant differences are observed. Analysis of the TAT task showed statistically significant differences in five patients out of eleven for the ampl* feature. The ampl* median values decrease in two patients passing from depression to euthymia, thus showing an opposite trend with respect to reading. Compared to the reading task fewer significant differences were

observed in hypomania for the task. When differences were found, the ampl* values in hypomania were higher.

TABLE 4.16: Median and median absolute deviation [mad] of Amplitude* estimated for bipolar patients. The symbols (∗ or +) indicate p-values <0.05 in Mann-Whitney U-test.

			READING			TAT		
	Mood state		Amplitude*			Amplitude*		
Subj.	day 1	day 2	day 1	day 2	p-value	day 1	day 2	p-value
A	Hyp.	Eut.	**0.12 [0.84]***	**-0.36 [0.63]***	**3.70E-03***	0.06 [0.72]	-0.05 [0.89]	3.04E-01
B	Hyp.	Eut.	**-0.34 [0.66]***	**-0.57 [0.42]***	**3.92E-02***	**-0.03 [0.88]+**	**-0.22 [0.77]+**	**4.26E-02+**
C	Hyp.	Eut.	**-0.45 [0.54]***	**-0.20 [0.80]***	**3.67E-02***	-0.33 [0.66]	-0.38 [0.61]	3.51E-01
D	Dep.	Eut.	-0.50 [0.49]	-0.59 [0.40]	9.38E-01	**-0.05 [0.94]+**	**-0.37 [0.62]+**	**1.41E-03+**
E	Dep.	Eut.	**-0.43 [0.57]***	**-0.04 [0.95]***	**2.43E-02***	**0.36 [0.55]+**	**0.03 [0.96]+**	**5.62E-03+**
F	Dep.	Hyp.	**0.32 [0.62]***	**0.47 [0.50]***	**3.99E-02***	**0.08 [0.72]+**	**0.28 [0.57]+**	**4.23E-03+**
G	Hyp.	Eut.	**0.44 [0.52]***	**0.27 [0.68]***	**2.71E-02***	-0.10 [0.89]	0.00 [0.87]	1.65E-01
H	Dep.	Eut.	0.15 [0.76]	0.06 [0.89]	6.69E-01	0.15 [0.68]	-0.21 [0.78]	4.58E-01
I	Dep.	Eut.	**-0.81 [0.18]***	**-0.25 [0.74]***	**1.55E-04***	-0.07 [0.72]	-0.04 [0.82]	8.01E-01
L	Dep.	Eut.	-0.07 [0.92]	0.12 [0.84]	4.17E-01	0.16 [0.83]	-0.09 [0.90]	4.78E-01
M	Mix.	Dep.	0.12 [0.71]	0.02 [0.62]	2.00E-01	**0.11 [0.64]+**	**-0.05 [0.76]+**	**2.61E-02+**

TABLE 4.17: Median and median absolute deviation [mad] of AbsNegSlope as estimated from bipolar patients. The symbols (∗ or +) indicate p-values <0.05 in Mann-Whitney U-test related to AbsNegSlope features.

			READING			TAT		
	Mood state		AbsNegSlope			AbsNegSlope		
Subj.	day 1	day 2	day 1	day 2	p-value	day 1	day 2	p-value
A	Hyp.	Eut.	**0.45 [0.31]***	**0.57 [0.36]***	**1.70E-03***	**0.61 [0.36]+**	**0.49 [0.32]+**	**9.60E-03+**
B	Hyp.	Eut.	**1.28 [0.97]***	**0.95 [0.72]***	**1.07E-02***	**0.57 [0.39]+**	**0.46 [0.31]+**	**3.29E-03+**
C	Hyp.	Eut.	**0.64 [0.34]***	**0.53 [0.31]***	**2.51E-02***	**0.44 [0.29]+**	**0.40 [0.23]+**	**2.22E-02+**
D	Dep.	Eut.	**0.32 [0.20]***	**0.39 [0.26]***	**1.56E-02***	0.38 [0.26]	0.42 [0.29]	5.68E-02
E	Dep.	Eut.	0.36 [0.18]	0.45 [0.32]	5.06E-02	0.43 [0.24]	0.47 [0.27]	2.00E-01
F	Dep.	Hyp.	**0.45 [0.26]***	**0.53 [0.30]***	**1.65E-02***	0.69 [0.60]	1.01 [0.70]	2.16E-01
G	Hyp.	Eut.	**0.69 [0.40]***	**0.61 [0.36]***	**2.10E-02***	**0.55 [0.34]+**	**0.63 [0.41]+**	**1.53E-02+**
H	Dep.	Eut.	**0.59 [0.36]***	**0.46 [0.31]***	**1.80E-02***	**0.46 [0.30]+**	**1.18 [0.67]+**	**6.90E-03+**
I	Dep.	Eut.	**0.65 [0.38]***	**0.49 [0.32]***	**5.39E-04***	0.40 [0.31]	0.48 [0.36]	6.46E-01
L	Dep.	Eut.	0.54 [0.38]	0.49 [0.36]	2.24E-01	**0.33 [0.22]+**	**0.48 [0.35]+**	**2.52E-02+**
M	Mix.	Dep.	0.64 [0.38]	0.58 [0.35]	9.45E-01	0.46 [0.30]	0.39 [0.27]	1.68E-01

Results pertaining to other features show statistically significant differences between mood states. However, no coherent direction of change in subjects experiencing the same mood swing could be observed. Moreover, the direction of change are not the same across the same tasks. In particular, dur* feature shows statistically significant differences in six patients out of eleven in reading, while, the TAT task differences were observed in two subjects only (data not shown). The analysis of the tilt* feature returns five significant p-values in reading, and three p-values in TAT task (data not shown). In six patients out of eleven, PosSlope shows statistically significant differences in reading and four differences in the TAT task (data not shown). The results pertaining to AbsNegSlope (Table 4.17) report eight significant differences in the reading task, and six statistically significant differences in TAT task. With regard to reading, in three patients out of four the AbsNegSlope is lower in the euthymic state than in hypomanic state. The same behaviour was found for the TAT task. However, the same direction of change

was found in both tasks only for patients B and C. The sumDer feature shows statistically significant differences in six patients out of eleven, regarding the reading, while no differences were observed regarding TAT task (data not shown). Finally, GlobalSlope in reading reports differences in seven out of eleven subjects (data not shown). In all cases, except two, GlobalSlope was closer to zero for euthymic subjects. In two subjects statistically significant differences were found only in one feature acquired during TAT. In particular, subject L showed statistically significant differences only in AbsNegSlope, while subject M only in ampl*. Group analyses are not performed here since the number of subjects is too small.

A pairwise comparison exploiting the hypomania-euthymia transition in four subjects and one exploiting the depression-euthymia transition in five subjects were performed (Table 4.18). Statistically significant differences were observed in AbsNegSlope concerning the transition between depression and euthymia during the performing of the TAT task.

TABLE 4.18: Results regarding paired inter-state analysis on bipolar data. In bold the statistically significant differences are highlighted. The significant differences are detected according to the critical values for Friedman's F_r reported in [19].

	Hyp. Vs. Eut.		**Dep. Vs. Eut.**	
	Reading	**TAT**	**Reading**	**TAT**
Amplitude*	>0.05	>0.05	>0.05	1,80
Duration*	>0.05	>0.05	>0.05	>0.05
Tilt*	>0.05	>0.05	>0.05	>0.05
PosSlope	>0.05	>0.05	>0.05	>0.05
AbsNegSlope	>0.05	>0.05	>0.05	**<0.05**
SumDer	>0.05	>0.05	>0.05	>0.05
GlobalSlope	>0.05	>0.05	>0.05	>0.05

4.5.1.3 Taylor's Extended Intonational Model: Feature Specificity and Healthy Control Subjects

Intra-subject analyses were carried out of data with the same label to check for the specificity of the features. To have good specificity, the feature should not show any statistically significant differences. Specificity was investigated by analysing data acquired on the same day from bipolar patients and the data acquired from healthy control subjects on different days. In Tables 4.19 and 4.20 the results are shown for bipolar patients recorded in the morning and in the afternoon at the first recording day (day 1). At that day, no statistically significant differences were found between Taylor-inspired features across all subjects (Table 4.19). At that day 2, a difference was found for subject B in ampl* and for subject E both in dur* and tilt* (data not shown). In Table 4.20 the results related to the second category of features estimated at day 1 from bipolar

subjects are summarized. SumDer and GlobalSlope did not show any statistically significant difference either at day 1 or at day 2 (data not shown). PosSlope revealed two statistically significant differences at day 1 and no differences at day 2. AbsNegSlope did not show any difference at day 1 and only one difference at day 2 for subject G (data not shown).

TABLE 4.19: Results at day 1 sessions concerning ampl*, dur* and tilt* for bipolar patients. Median and Mad (in square brackets) values are shown. No statistically significant differences were found.

Subj.	ampl*		dur*		tilt*	
A	0.05 [0.86]	0.12 [0.84]	-0.50 [0.45]	-0.40 [0.59]	-0.42 [0.57]	-0.34 [0.65]
B	-0.57 [0.43]	-0.34 [0.66]	-0.69 [0.30]	-0.66 [0.33]	-0.71 [0.28]	-0.62 [0.37]
C	-0.45 [0.54]	-0.40 [0.25]	-0.66 [0.33]	-0.63 [0.36]	-0.67 [0.32]	-0.67 [0.32]
D	-0.50 [0.49]	-0.59 [0.40]	-0.50 [0.46]	-0.50 [0.45]	-0.31 [0.68]	-0.47 [0.52]
E	-0.43 [0.57]	-0.40 [0.95]	-0.75 [0.25]	-0.77 [0.22]	-0.75 [0.24]	-0.78 [0.21]
F	0.40 [0.54]	0.32 [0.62]	-0.33 [0.66]	-0.30 [0.69]	-0.05 [0.94]	-0.09 [0.90]
G	0.38 [0.57]	0.44 [0.52]	-0.33 [0.66]	-0.36 [0.63]	-0.16 [0.83]	-0.16 [0.83]

TABLE 4.20: Results at day 1 concerning SumDer, GlobalSlope, PosSlope and AbsNegSlope as estimated for bipolar patients. The symbol $*$ indicates p-values <0.05 (Mann-Whitney U-test).

Subj.	SumDer		GlobalSlope		PosSlope		AbsNegSlope	
A	0.16 [0.36]	0.17 [0.34]	-0.05 [0.66]	-0.01 [0.69]	0.54 [0.32]	0.58 [0.35]	0.44 [0.29]	0.45 [0.31]
B	0.10 [0.61]	0.17 [0.71]	-0.40 [1.42]	-0.44 [1.50]	0.85 [0.57]	0.83 [0.52]	1.00 [0.75]	1.28 [0.97]
C	0.12 [0.44]	0.13 [0.40]	-0.27 [0.95]	-0.25 [0.97]	0.70 [0.43]	0.75 [0.39]	0.64 [0.34]	0.60 [0.36]
D	0.18 [0.30]	0.16 [0.22]	-0.02 [0.64]	-0.03 [0.57]	**0.52 [0.23]***	**0.38 [0.19]***	0.32 [0.20]	0.39 [0.24]
E	0.07 [0.27]	0.15 [0.33]	-0.22 [0.62]	-0.24 [0.62]	**0.45 [0.26]***	**0.59 [0.31]***	0.36 [0.18]	0.36 [0.20]
F	0.33 [0.42]	0.40 [0.48]	0.15 [0.71]	0.17 [0.66]	0.78 [0.43]	0.79 [0.40]	0.47 [0.30]	0.45 [0.26]
G	0.47 [0.70]	0.30 [0.63]	0.13 [0.92]	0.20 [0.88]	1.11 [0.63]	0.99 [0.60]	0.73 [0.45]	0.69 [0.40]

TABLE 4.21: Results regarding paired inter-state analysis on bipolar data. In bold the statistically significant p-values are highlighted.

	Reading	**TAT**
Amplitude*	8.96E-02	5.27E-01
Duration*	8.08E-01	5.27E-01
Tilt*	2.25E-01	5.27E-01
PosSlope	8.08E-01	5.27E-01
AbsNegSlope	8.08E-01	1.00E+00
SumDer	8.08E-01	1.00E+00
GlobalSlope	2.25E-01	5.27E-01

Concerning the Healthy Control Subjects Database (data not shown), while reading ampl*, dur* tilt* and GlobalSlope did not show any statistically significant differences between different days. PosSlope, AbsNegSlope and SumDer showed statistically significant differences in 3, 4 and 2 subjects out of 18 respectively. While TAT commenting, ampl*, tilt*, AbsNegSlope and GlobalSlope did not show any statistically significant differences, while dur* and AbsNegSlope reported a significant difference in 1 out of 10 subjects, PosSlope in 3 out of 10, and SumDer in 2 subjects out of 10.

At the end no statistically significant differences were found between the global average features extracted from the audio recordings acquired in the two daily experimental sessions (Table 4.21).

4.5.2 Spectral analysis of intonational contour

4.5.2.1 Spectral analysis of intonational contour: Bipolar Data

The proposed features showed a similar behaviour across all subjects. Specifically, F_{peak} was always lower than F_{median} thus resulting in a negative *Slope* cue and in a $Ratio_{peak}$ that was always higher than $Ratio_{median}$.

Each bipolar patient in an euthymic state in one of the three recording days was selected for this study. By exploring Tables 3.1 and 3.2 it is possible to select the subjects that can be used for paired and independent data tests. In general, the reading task in bipolar patients took about 4 minutes. Analysis of paired data (Table 4.22) showed statistically significant differences between hypomania and euthymia states (patients A, B, C and G) for A_{peak}, F_{peak}, $Ratio_{peak}$ and their *Slope*.

TABLE 4.22: Bipolar patients: p-values. In bold the statistically significant differences are highlighted. The significant differences are detected according to the critical values for Friedman's F_r reported in [19].

	F_{median}	A_{median}	F_{peak}	A_{peak}	*Slope*	$Ratio_{peak}$	$Ratio_{median}$
Hyp vs. Eut	>0.05	>0.05	**<0.05**	**<0.05**	**<0.05**	**<0.05**	>0.05
Dep vs. Eut	**<0.05**	>0.05	>0.05	**<0.05**	**<0.05**	>0.05	>0.05

In all subjects but one (patient C), F_{peak} (Figure 4.22) was lower in the hypomanic state, while for all the subjects A_{peak} (Figure 4.23) and $Ratio_{peak}$ (Figure 4.24) was higher in the hypomanic state compared to the euthymic one. Opposite trends were observed for the *Slope* feature (Figure 4.25).

Moreover, further analysis of paired data (patients B, D, F, H, I, L and M) showed that differences between depression and euthymia states were statistically significant for F_{median}, A_{peak} and *Slope*. For all subjects, F_{median} (Figure 4.26) and Slope (Figure 4.28) were lower in the depressed state, while A_{peak} (Figure 4.27) was higher.

Comparisons carried out via the Mann-Whitney U-test on unpaired normalized data between depression and hypomania and depression and euthymia, showed statistically significant differences for F_{median} (Figure 4.29) and *Slope*. In addition A_{peak} reported significant differences just for depression-hypomania. F_{median} and *Slope* for the depressed state were lower with respect to the other mood states, while A_{peak} was higher. Without normalization, the features did not show any statistically significant differences. With

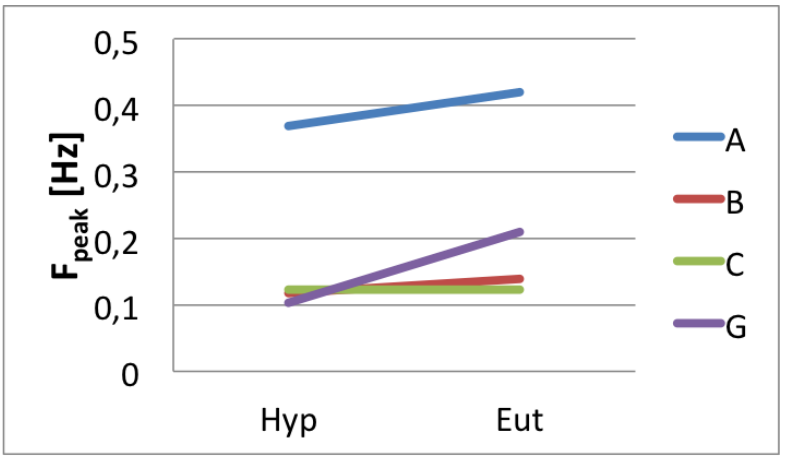

FIGURE 4.22: F_{peak} trends in patients passing from hypomania to euthymia.

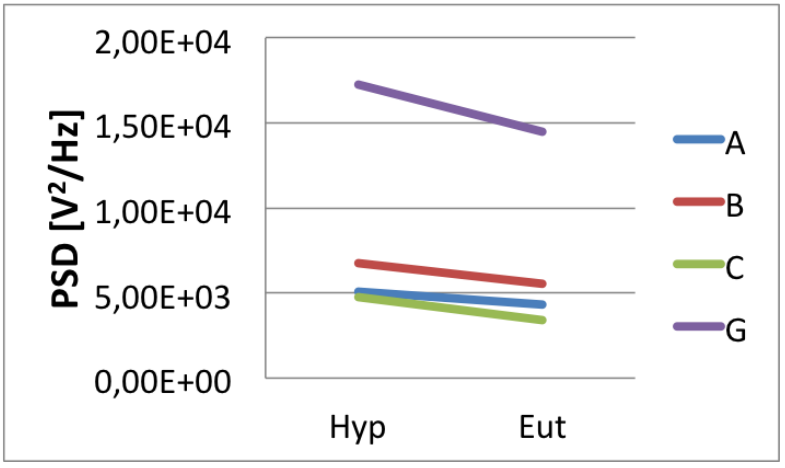

FIGURE 4.23: A_{peak} trends in patients passing from hypomania to euthymia.

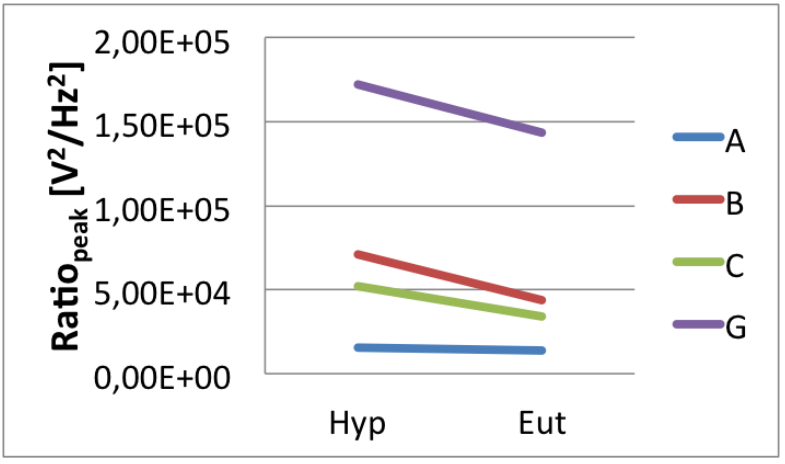

FIGURE 4.24: $Ratio_{peak}$ trends in patients passing from hypomania to euthymia.

a view to the Mann-Whitney U-test, data groups were formed with features of patients in different mood states.

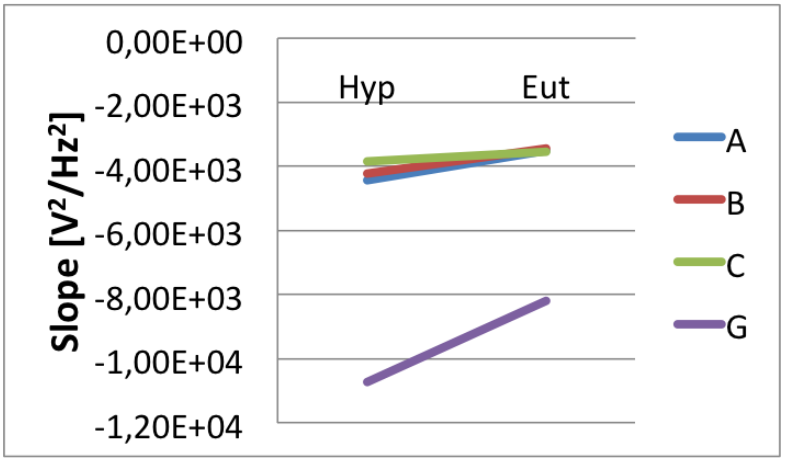

FIGURE 4.25: *Slope* trends in patients passing from hypomania to euthymia.

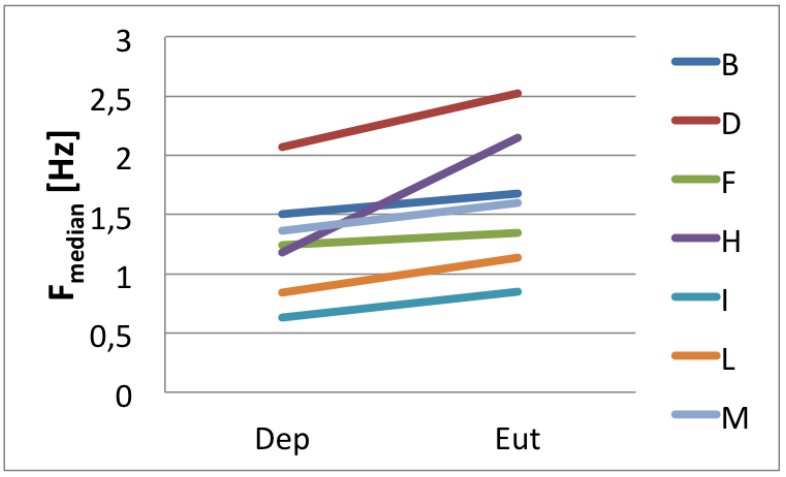

FIGURE 4.26: F_{median} trends in patients passing from depression to euthymia.

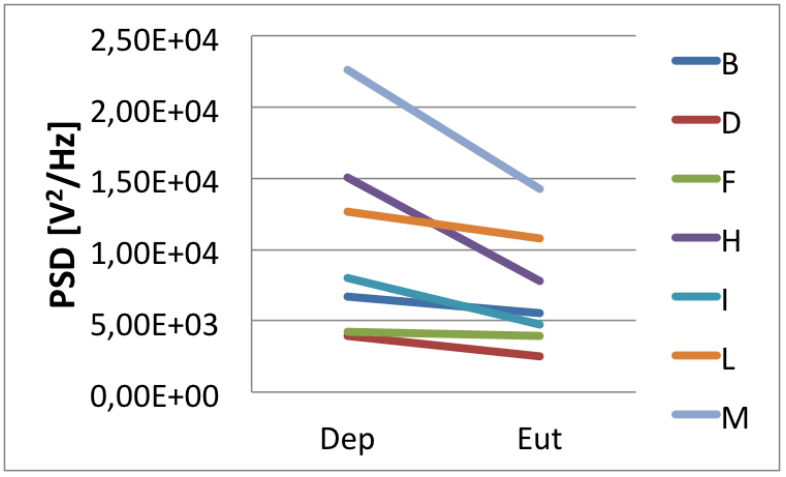

FIGURE 4.27: A_{peak} trends in patients passing from depression to euthymia.

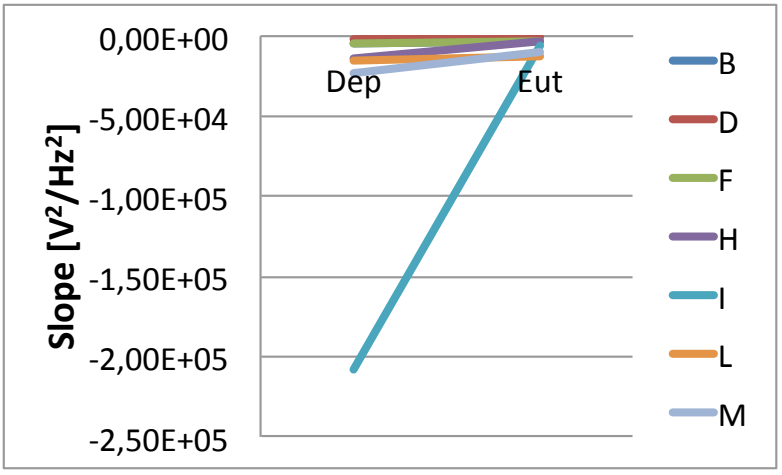

FIGURE 4.28: *Slope* trends in patients passing from depression to euthymia.

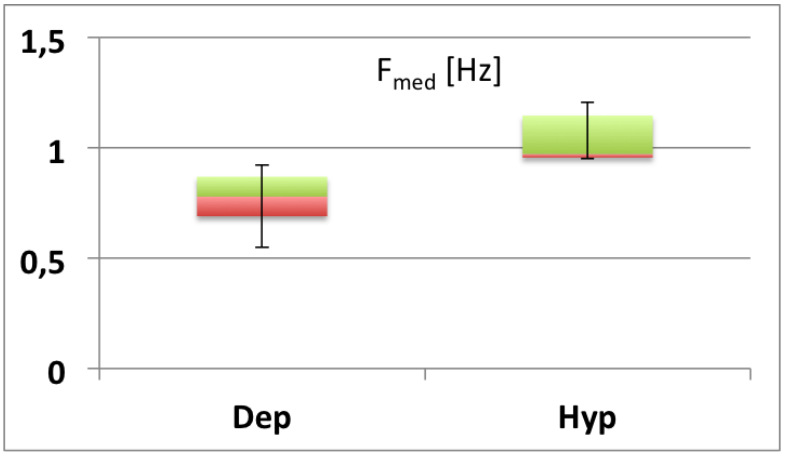

FIGURE 4.29: Boxplot of F_{median} in patients passing from depression to hypomania. F_{median} values are normalized with respect the corresponding values in euthymic state.

4.5.2.2 Spectral analysis of intonational contour: Features Specificity

Analysis of the data recorded for Healthy Control Subjects did not return statistically significant differences between features obtained from audio samples acquired at two different days. In Table 4.23 the corresponding p-values are reported, while in Figure 4.31 the F_{median} trends in healthy control subjects are displayed

TABLE 4.23: Healthy control subjects: p-values.

F_{median}	A_{median}	F_{peak}	A_{peak}	*Slope*	$Ratio_{peak}$	$Ratio_{median}$
7.39E-01	1.00E+00	2.06E-01	5.27E-01	5.27E-01	1.00E+00	5.27E-01

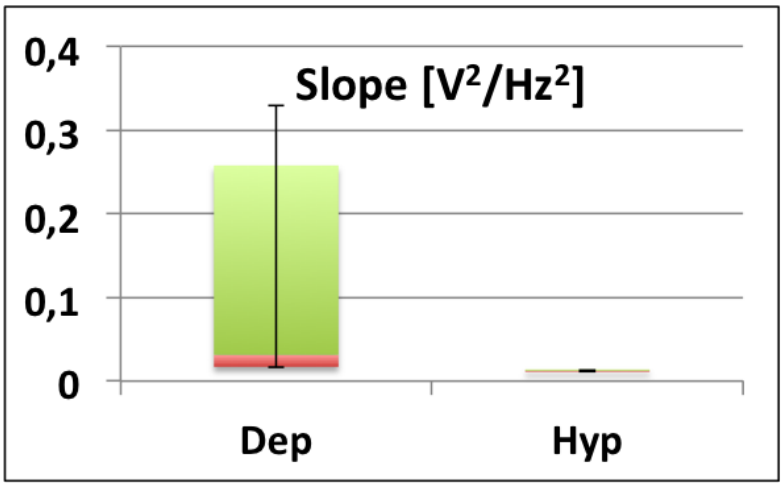

FIGURE 4.30: Boxplot of *Slope* in patients passing from depression to hypomania. *Slope* values are normalized with respect the corresponding values in euthymic state.

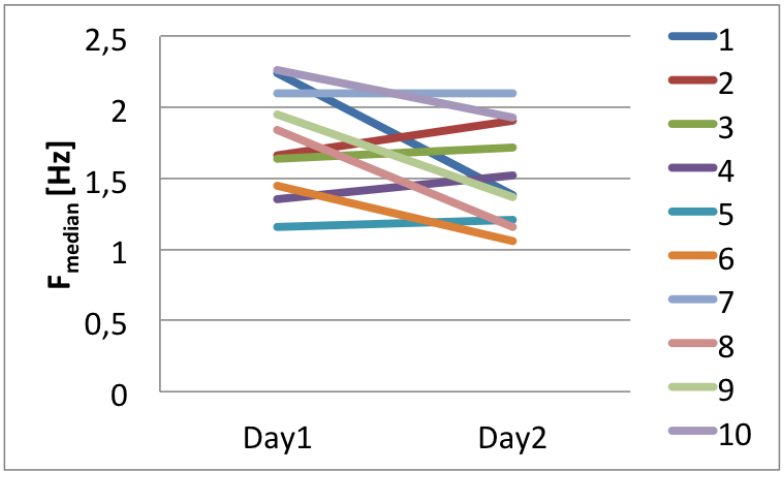

FIGURE 4.31: F_{median} trends in healthy control subjects.

4.6 Voice quality study

4.6.1 Voice quality: Analysis of F0 correction on synthetic data

A set of 6 synthetic voice samples was synthesized with mean-F0 equal to 100 and 150 Hz, and F0 jitter equal to 0.04, 1.2 and 2.0 %. In Figure 4.32 the LTAS profiles obtained on vowels at different F0 are compared. The results of the F0 corrected approach and the conventional approach are shown in different graphs. A percent change between the $150Hz$ and the $100Hz$ vowel spectra equal to -5.7% was obtained using the F0-corrected algorithm. The conventional approach resulted in a percent change equal to 28.9%.

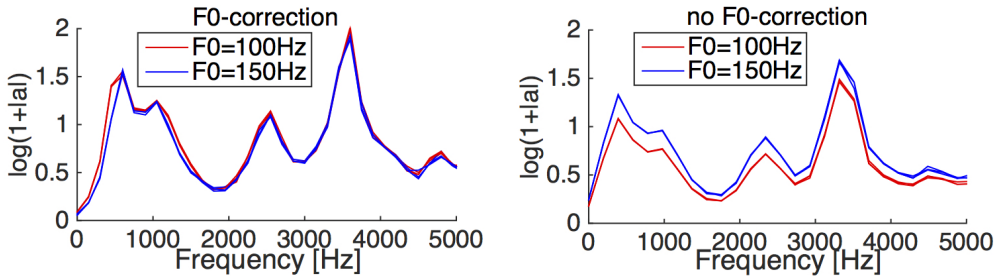

FIGURE 4.32: Differences in LTAS: F0-correction (left), and the conventional method (right). Vowels were synthesized with F0=150Hz(blue) and F0=100Hz(red).

4.6.2 Voice quality features: Emotion database

As previously discussed, meanF0 is able to detect statistically significant differences among different levels of arousal.

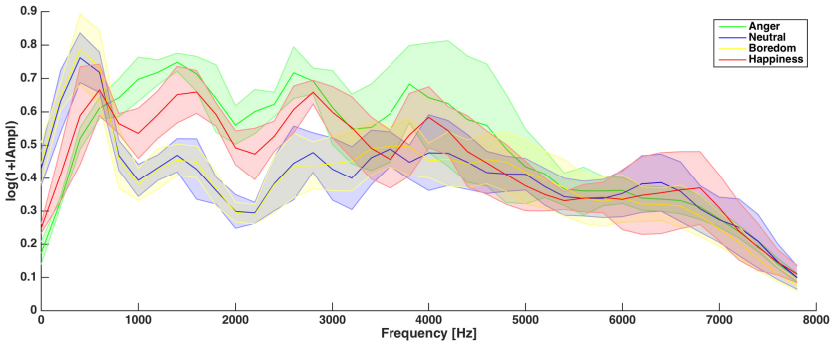

FIGURE 4.33: Conventional LTAS: Median and median absolute deviation (MAD) of LTAS for each emotion and for all the speakers.

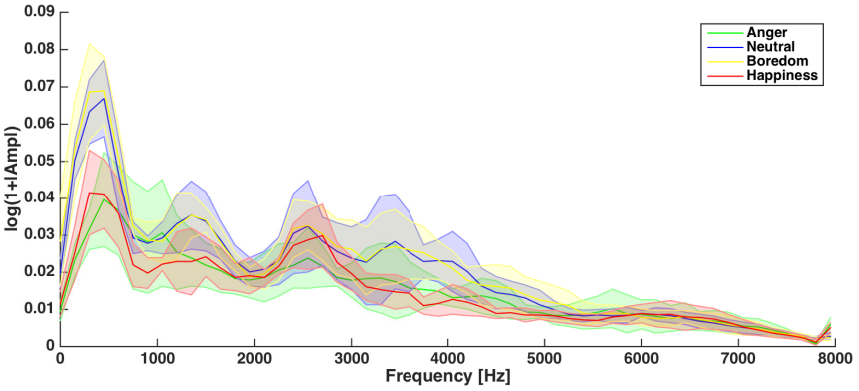

FIGURE 4.34: F0-corrected LTAS: Median and median absolute deviation (MAD) of LTAS for each emotion and for all the speakers.

Both algorithms were used to investigate possible statistically significant differences in frequency content related to emotional states in the German emotional database [137]. Grouped analyses revealed statistically significant differences among the four emotional states in the 0-600Hz and 800-3200Hz sub-bands when using the conventional LTAS

approach. In the first sub-band, lower arousal shows higher LTAS amplitudes, while, on the contrary, in the second band lower arousal states shows lower LTAS amplitudes. Figure 4.33 reports the median and the median absolute deviation of the LTAS for each emotion, which confirm the statistical analysis.

When using the F0-corrected LTAS algorithm, statistically significant differences were found among emotional states in the 0-600Hz and 1500-1650Hz sub-bands. In both sub-bands, lower arousal emotions show higher LTAS amplitudes.

4.6.3 Voice quality features: Healthy Control Subjects and Bipolar patients

Statistical analyses on LTAS extracted from Healthy Control Subjects' audio signals by means of both algorithms did not reveal any statistically significant differences. Friedman's test was used to analyse differences in frequency components, and the estimated statistics were below the critical 0.05 value.

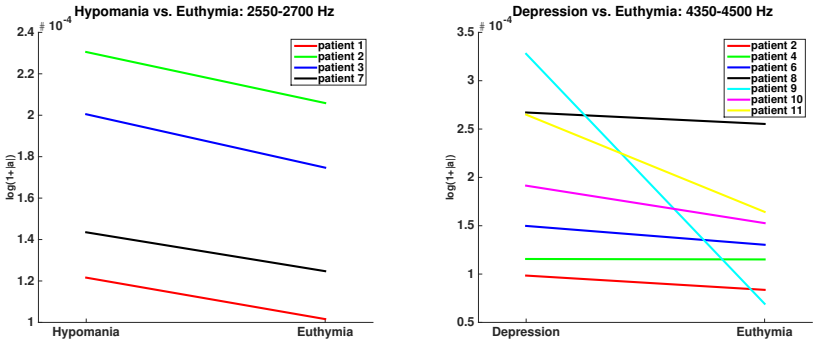

FIGURE 4.35: F0-corrected LTAS: trend of frequency content regarding different bins in the two pairwise tests.

Regarding bipolar patients, both algorithms showed similar results on pairwise statistical tests on reading task between hypomania and euthymia states (four patients). The conventional LTAS algorithm showed statistically significant differences in three sub-bands: 2600-2800Hz, 5800-6000Hz and 7200-7800Hz. Similarly, the F0-corrected algorithm highlighted significant differences in 2400-2700Hz, 5850-6000Hz and 7050-7800Hz sub-bands. In both approaches, LTAS was lower in the first sub-band in the euthymic state. The opposite behaviour was observed in the other sub-bands. The comparison between depression and euthymia was performed on seven patients. In this case, the two investigated algorithms reported dissimilar results on pairwise statistical tests. In fact, the conventional algorithm reported a higher amplitude in euthymia in the 2200-2800Hz sub-band, while the F0-corrected algorithm resulted in a lower amplitude in euthymia

in 4350-4500Hz sub-band. In Figure 4.35 the trends of F0-corrected LTAS obtained in both tests are shown. The results regarding one frequency bin are shown for each test.

Pairwise analysis of mean-F0 shows how changes between hypomania and euthymia states are not statistically significant, while mean-F0 was found to be lower in depressed state with respect to the euthymic one.

Concerning TAT task, conventional LTAS algorithm showed statistically significant differences only in hypomania-euthyia transition in the sub-band 5600-6400Hz. Similarly, F0-corrected algorithm highlighted significant differences in 5700-5850Hz sub-band only by comparing hypomania and euthymia states. Both methods showed no statistically significant differences in the depression-euthymia comparisons during TAT task. Moreover, no statistically significant differences were observed by analysing meanF0 in relation to the TAT task in both mood comparisons.

Chapter 5

Discussion and conclusion[1]

A study of the speech signal in an emotion and mood recognition/characterization framework was performed at different levels of description. Starting from a micro-prosodic study, higher level phenomena were investigated later. Initial investigations were focussed on small changes of the glottal cycle related to emotion and mood. Global prosodic and vocal quality studies were conducted later. The former reported overall intonational behaviour. The latter not being exclusively related to the glottal source, but also to the vocal tract, reported vocal timbre overall in syllable nuclei.

5.1 VAD

Every step in feature extraction was tested and evaluated by means of analysis of real and synthetic audio sample. Concerning voice activity detection, results on synthetic data revealed that both algorithms, the proposed one and the one used as benchmark, showed similar performances. Different voiced-unvoiced transients and F0 contours were simulated and analysed to test VAD algorithms. Each of the 72 audio samples involved 2 vowels and an unvoiced segment. The weak point of the VAD methods was the voiced-unvoiced sound transition lengths. The longer the transition, the larger the error was. The proposed VAD algorithm showed slightly better performance in terms of error detection. This test was performed after the optimization of the analysis parameters. Optimization was performed by means of the CMU Arctic Database. This database, involving both audio and EGG signals, enabled us to detect, by using real data, the exact time instants of glottal activity. Therefore parameters were fixed to obtain a higher specificity with respect to sensitivity, without the loss of great amount of relevant data. Concerning the criteria for the VAD parameters selection, a specificity higher than

[1]Part of this Chapter has been already published in [1–5].

sensitivity was chosen to reduce the number of unvoiced segments that are recognized as voiced, and hence to reduce the probability to analyse incorrectly labelled unvoiced sounds. A specificity higher than 0.90, and a sensitivity higher than 0.80 were selected.

5.2 F0-estimation

Concerning the F0-estimation algorithm tests were performed on synthetic data. The study exploits Camacho's SWIPE' algorithm for F0 estimation. Camacho in [16] compared his method with others, reporting good performance. Evanini confirmed these results in [316]. In [300], the SWIPE' algorithm was used to estimate F0 and jitter on voiced segments, and its performance was compared with those of the Simplified Inverse Filter Tracking algorithm (SIFT) [317]. The performances of the two approaches were similar concerning average F0 on each voiced segment. But, SWIPE' outperformed the SIFT with regard to jitter estimation. In this study, 512 audio signals were synthesized and used to test the method developed. The results showed a median error lower than 0.3%, highlighting a good capability in the proper estimation of voice fundamental frequency.

Results obtained with the CMU Arctic Database, confirmed that the proposed approach can be used to estimate local mean F0 values reliably. A high correlation between local F0 standard deviations estimated with the proposed approach and the EGG signal was obtained as well. The correlation between estimated frame-to-frame jitter and the benchmark value obtained by the EGG signal was lower, but statistically significant. The proposed approach estimates one F0 value within a time window of four glottal cycles. This enables reliable results in noise, but causes a systematic jitter underestimation [300]. Moreover, the analysis showed that meanF0 is robust with regard to noise. StdF0 and LpJ estimates remain accurate in several noisy conditions. LpJ estimates were degraded at $5dB$ SNR. The results also demonstrate that echo strongly affects features reporting F0 changes. This suggests a possible issue with indoor recordings.

5.3 Detection of emotional states

Concerning the analysis of emotion data, one may say that the more the subjects are aroused, the more their speech features exhibit differences with low arousal states. These results seem to be in agreement with Pakosz, who asserted that intonation can only carry information about the level of emotional arousal [318]. Moreover, Banse and Scherer showed how arousal has a powerful effect on vocal expressions often hiding the effects of valence or potency/control [319].

The emotion database we took into account is a collection of sentences spoken by actors who were "playing" different emotions, while the actors' actual mood is unknown. Vogt and André [105] showed that feature sets recognizing different emotions in acted and spontaneous speech overlapped only partially. Bänziger and Scherer [106] defended in a detailed way the prudent use of acted sentences in the study of emotion. The difficulties to record different and often rare emotional states from the same subjects, and to assess each emotional state might assign acted speeches datasets important role in this field. Schuller et al. [107] asserted that acted corpora have two disadvantages: the first is that acting emotions is different from producing "spontaneous" emotions [108] and secondly, the prompted types of emotions are not the same as those in realistic scenarios. So while the acquisition of realistic corpora is envisaged, using acted corpora could be convenient for benchmarking, even if the relationship between the results coming from the two kind of dataset is unclear [109]. Here, the emotion dataset was important to evaluate the capability of the implemented algorithm to extract prosodic features, confirming the algorithm's capability to estimate subjects' mood state.

Therefore, in the present study, the adequacy of the analysis for bipolar patients speech is not inferred from the results on the emotion database. The pathophysiological factors influencing speech in bipolar disorders could lead to completely different phenomena linking subject mood states to voice production.

5.4 Data

Concerning the analysis on bipolar data, a study involving a larger number of subjects would reveal the behaviour of the proposed features more clearly. A subject-dependent behaviour of the features can indeed not be excluded. Subject anxiety, for instance, could be a psychological dimension that may be taken into account in further research. In [215], observed differences were hypothesized to be caused by anxiety. As a possible confirmation of this hypothesis, in [320] the authors suggested that some vocal parameters, for example F0, can be used as objective markers of Social Anxiety. Anxiety level could also be a factor affecting specificity, even though we cannot exclude the relevance of other unobserved factors. In particular, as regards the features extracted from bipolar patients during double recording sessions, uncontrolled factors could include boredom or fatigue during a full day visit to the clinic. A further confounding phenomenon could be related to the subjects' familiarity with the text to be read. Agitated and retarded kinds of depression or even hypomanic and hypermanic kinds of mania or mixed states [321] could also explain the incoherent trends that some features displayed. The bipolar patients enrolled in this study did not show severe symptoms, while in other studies severe

depression has been included. The lack of severe symptoms as well as the small number of participants may limit the possibility detecting statistically significant changes in the investigated features. Another possible limitations could be the patients' willingness to please the clinicians and showing themselves healthy. Patients' behaviour could be a confounding factor in this study. Pharmaceutical treatment could be also a confounding factor. Often drugs cause a reduction in the amount of patients' saliva. Indeed, some antidepressants or psychotropic drugs induce xerostomia (i.e., dry mouth) as a frequent side effect [322, 323]. Such a reduction could produce some modification in the acoustic properties of the vocal tract that are not related to the disease, but to its medication.

Moreover, no statistical analyses of inter-task results were carried out. However, the directions of feature changes sometimes are not the same across the two tasks. A task dependent behaviour of F0 has also been observed by Horwitz et al. [262] in depressed patients where the correlations between F0 and a score of depression was investigated. The two tasks were the reading of a text and a sample of TAT speech. The differences, that are reported here, might be explained by taking into account the differences between the two tasks. The description of the images during TAT, involves complex brain processes since they require an interpretation of the images.

The database included speakers with different native languages, i.e. French and Italian. This can have a minor effect when intra-subjects analysis or inter-state analysis for paired data are performed, but could have a major effect when group-level analyses are performed.

5.5 Vocal features

Both intra-subject and group analyses of vocal features obtained from the German Emotional Database demonstrated the capability of a subset of features to report statistically significant differences between audio samples related to different arousal levels, meanF0 and stdF0 in particular. More precisely, statistically significant differences were found for these features estimated from audio samples conveying anger/happiness versus neutral/boredom. No differences were observed when investigating vocal jitter. Jitter only reported between boredom and happiness. In some subjects, statistically significant differences were observed between high arousal emotional states (anger vs. happiness) or between low arousal emotional states (neutral vs. boredom). It was not possible to obtain such a result at the group level.

Results on bipolar patients were the following. Intra-subject analysis revealed possible differences in speech samples acquired in the same task category, but between the different sessions reporting different mood states. The small number of subjects enrolled limits the generalizability of the results, but some observations can be made.

The analysis of F0 revealed statistically significant differences between different mood states in all subjects but one, in one of the two tasks considered. When the direction of change was observed in both tasks, it was found to be the same in all the subjects but one. The observed changes are not always consistent across subjects. MeanF0 was lower in the depressed state in two subjects (H and I) while the opposite behavior was observed for other two subjects (F and L). The former trend is more frequently reported in the literature, but the latter has been associated with anxiety [215]. This observation suggests the need for improving the characterization of the subjects' psychological status and taking into account anxiety to clarify possible interactions between mood and anxiety. When statistically significant differences were observed in hypomania-euthymia transitions, higher feature values were detected in the hypomanic state in all cases but one. In fact, in subject G one observes a lower feature value in hypomania in the TAT task. This subject is the one who reported incoherent change directions between the two tasks.

When statistically significant differences were observed, StdF0 was found to be lower in the euthymic state with respect to the hypomanic one in all cases but one. Subject G showed the opposite trend in recordings related to the TAT task. This subject reported different feature trends between the two tasks also for stdF0.

Concerning the reading task, LpJ was found to be higher in the hypomanic state in two out of three subjects who showed statistically significant differences. The same trend was observable in two out of three subjects who showed statistically significant differences regarding the TAT task. Incoherent trends were observed in subjects showing statistically significant differences between LpJ extracted in the depressed and euthymic states. Subject G show different feature trends between the two tasks also in LpJ.

The results on healthy subjects, using paired tests, do not highlight any differences due to the repetition of the tasks at different days. This result only partially addresses the specificity issue, which would be better characterized by analysing patients experiencing the same mood in different days.

Interestingly, the analysis of the reading task highlighted that meanF0 decreases when passing from the hypomanic to the euthymic state. Statistically significant differences were found in LpJ in subjects switching from depression to euthymia. In this case, the sign of the change of LpJ was discordant with observations reported in the literature

[222]. However, in that study, the changes were observed in subjects with severe symptoms. The lack of significant differences in stdF0 in the case of depression-euthymic transitions seem also to contradict literature results. However, our results might be related to the choice of estimating stdF0 at syllable level.

Although the small number of patients limits the generalizability of our results, they confirm that the speech task has a relevant role since it influences speech feature changes [179]. Differences in average F0 (meanF0) were observed only in the text reading task, while LpJ changes were observed only in the TAT commenting task. In conclusion, the analysis of vocal features in bipolar patients could provide information regarding mood state changes. The speech task has been shown to be relevant and deserves further investigation.

5.6 Taylor's Extended Intonational Model

In this study two categories of prosodic features are used. The first is inspired by Taylor's Tilt Intonational model. The features used here are only morphologically equivalent to Taylor's. In the tilt model intonational events are taken into account, while in this study the features were estimated for all voiced vowel nuclei of syllables. The detection of intonational events relies on the ability of the human labeller and requires training an automatic classifier starting from hand labelled sentences. The here proposed approach is simpler and completely automatic. The second category of features, differently from the one proposed by Taylor, reports the speed of variation of F0 in the F0 contour. The here proposed features set supplies therefore an higher information content.

Analysis on the emotion speech database demonstrated that the proposed features enable highlighting significant differences among different emotional speech recordings. Such differences were observed both in subject and in group analyses. In particular, some features have been shown to be capable of grouping emotions by arousal level.

Intra-subject analyses on bipolar patients have shown that prosodic features have a good specificity. In almost every comparison, between features extracted from different acquisitions and labelled with the same mood state, no statistically significant differences were found. To test for specificity in bipolar patients, double recording sessions were performed in the same day. Good results were found by analysing data acquired from healthy subjects at different days. In particular, Taylor-inspired features and GlobalSlope demonstrated very high specificity. The remaining features showed a good specificity with regard to reading, while worse results were found for TAT recordings. As

a result, statistically significant differences between features reported a different mood states.

Overall, this study shows that the direction of the change is not coherent across subjects. Only ampl* seems to have a coherent behaviour across subjects. In particular, when statistically significant differences were found, the ampl* values extracted from recordings related to hypomania are higher than the other ones in every subject, but one, irrespective of the task. Noticeably, when one of the two states was the hypomanic state, a difference was always found regarding the reading task. The same behaviour was not observed when analysing TAT recordings.

5.7 Spectral analysis of the intonational contour

In this study, a spectral analysis of the F0-contours is carried out. Conventionally, the F0-contour is studied in the time domain. An analysis in the frequency domain might provide a compact description of the F0-related prosodic information. Specifically, the analysed features report the shape of the F0 spectrum profile. Since F0 was set to 0 Hz in silent segments, the features summarize the contribution of rhythm as well as of intonation. Specifically, results depend not only on syllabic rhythm (4 Hz typically), but also on pauses between words and sentences.

The statistical analysis was performed on bipolar patients experiencing different mood states. Moreover such a method was also applied on an Healthy Control Subjects Database. Statistically significant differences were found between features across different mood states. Interestingly, the proposed features showed a good specificity, whereas they were similar for control subjects. Notwithstanding the small number of patients who have been analysed, the results may be relevant because coherent feature trends have been detected in patients across mood states. Due to the sample size, it was not possible to perform any statistical test on paired data with regard the comparison between depression and hypomania. The comparison of euthymia and depression showed that $_{Apeak}$ increases and F_{median} decreases in the depressed state with respect to the euthymic stated. Since F_{peak} is lower than F_{median}, this behavior suggests a higher contribution at lower frequencies in the depressed state. The paired analysis of hypomanic and euthymic states revealed a significant decrease of F_{peak} and an increase of A_{peak} in the former state, while no relevant change of F_{median} was observed. Moreover, a decrease of the Slope cue was found. These results show a behavior of the analysed features that may possibly differentiate hypomania from depression. Preliminary results on independent samples seem to confirm this hypothesis. In fact, a decrease of F_{median} and Slope was reported in depressed with respect to hypomanic patients, thus indicating a

higher contribution at lower frequencies of the F0 profile spectra in the former subjects. The significant results, obtained from the statistical tests on independent samples, were reached after normalizing the feature values by the corresponding value in the euthymic state. This normalization was performed under the hypothesis that euthymia represents the emotional point of reference since it is characterized by the absence of relevant symptoms.

The choice of the text may play a crucial role both with regard to content and the structure. A specific content might elicit an emotional response. Moreover, since the reading rhythm is one of the parameters under study, it is important to use text with similar or equal lexical structure. In this study, a neutral text was adopted, i.e. "The universal declaration of human rights" for the different recording sessions.

5.8 Voice quality

This study aims at investigating voice quality in patients suffering from bipolar disease. According to Laver's model [75], phonation or vocal tract configurations departing from neutral settings causes the emission of a coloured voice. Settings can be divided in two groups: settings of the larynx and settings of the supralaryngeal vocal tract. In particular the laryngeal ones define phonation types.

A F0-corrected LTAS algorithm is discussed and tested on synthetic audio samples. The proposed algorithm takes into account glottal closure instants to set, cycle by cycle, the frame length and position of the sliding window instead of fixing these globally. Simulations confirm the effectiveness of the proposed F0-correction. LTAS seem to be influenced by F0 also at higher frequencies. This is also in accordance with a study conducted by Cleveland, Sundberg and Stone [324], who asserted that LTAS reflects the contribution of both the glottal source and the vocal tract to voice quality.

Analysis of healthy control subjects by both investigated algorithms does not show any statistically significant differences between different recording days.

Comparing the two proposed algorithms on the German Emotional Database, statistically significant differences were observed in different sub-bands. Conventional LTAS reported some statistically significant differences also at a higher frequency than the ones reported by the F0-corrected LTAS method. The differences were observed in accordance with the level of arousal. Differences between high arousal level emotions, i.e. anger and happiness, and low arousal level emotional states, i.e. neutral and boredom, were detected. In particular low frequency components showed higher amplitude in the lower arousal emotions.

Bipolar patients were labelled differently with regard to mood at each acquisition day. LTAS show statistically significant differences in some frequency intervals. These are found in both pairwise comparisons of hypomanic/euthymic states and depressed/euthymic states. Some differences may depend on modifications of formants due to the mood state. In fact, a decrease in the second and third formant, in depressed with respect to healthy subjects, has been described in [213]. The opposite trend is reported in [221].

Concerning the reading task, F0-correction does not influence the results when hypomania and euthymia are compared. On the contrary the two investigated algorithms seem to provide different results when comparing depression and euthymia. Such differences may depend on the statistically significant variation of F0 found out between depression and euthymia. In this study, F0 estimation was obtained by using the same VAD algorithm used to estimate LTAS from every voiced segment. More specifically, the method based on the autocorrelation function and the signal energy was here implemented. Coherent results were obtained by analysing audio samples related to the TAT task via the two LTAS approaches. Anyway, in this case, no statistically significant variations of F0 were observed.

Since multiple tests within different frequency bins were performed, an adjustment for multiple comparisons might be performed. In this study, an adjust for multiple comparisons was not performed. In fact, given the small sample size it was not possible to have an exact p value for the Friedman F statistics, and no randomization tests could be reliably performed. Even if the limited number of enrolled patients does not allow to generalize results, the proposed methods have been shown to have good specificity, because statistically significant differences are related to different mood states. In fact, no differences were found between healthy control subjects recorded at different days. This result confirms the significance of other results obtained in bipolar patients.

Bibliography

[1] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanata, and E.P. Scilingo. Speech analysis for mood state characterization in bipolar patients. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2104–2107. IEEE, 2012.

[2] A Guidi, N Vanello, G Bertschy, C Gentili, L Landini, and EP Scilingo. An automatic method for the analysis of pitch profile in bipolar patients. *Models and Analysis of Vocal Emissions for Biomedical Applications*, page 231, 2013.

[3] A Guidi, N Vanello, G Bertschy, C Gentili, L Landini, and EP Scilingo. Automatic analysis of speech f0 contour for the characterization of mood changes in bipolar patients. *Biomedical Signal Processing and Control*, 17:29–37, 2015.

[4] A Guidi, J Schoentgen, G Bertschy, C Gentili, EP Scilingo, and N Vanello. A spectral analysis of f0-contours in bipolar patients. *Models and Analysis of Vocal Emissions for Biomedical Applications*, page 131, 2015.

[5] A Guidi, J Schoentgen, G Bertschy, C Gentili, L Landini, EP Scilingo, and N Vanello. Voice quality in patients suffering from bipolar disease. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 6106–6109. IEEE, 2015.

[6] Lawrence R Rabiner and Ronald W Schafer. Introduction to digital speech processing. *Foundations and trends in signal processing*, 1(1):1–194, 2007.

[7] Jody Kreiman and Diana Sidtis. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons, 2011.

[8] 07 2015. URL http://www.britishvoiceassociation.org.uk/voice-information_vocal-nodules.htm.

[9] Frank H Netter. *Atlas of human anatomy*. Elsevier Health Sciences, 2014.

[10] July 2015. URL http://voicefoundation.org/wp-content/uploads/2013/10/vfscarring_normal_large.jpg.

[11] James L Flanagan. *Speech analysis: Synthesis and perception..* Springer-Verlag, 1972.

[12] JRJG Proakis, JR Deller, and JHL Hansen. Discrete-time processing of speech signals. *New York, Macrnillan Pub. Co*, 1993.

[13] Gordon E Peterson and Harold L Barney. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184, 1952.

[14] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice hall, 1993.

[15] Anthony Holbrook and Grant Fairbanks. Diphthong formants and their movements. *Journal of Speech, Language, and Hearing Research*, 5(1):38–58, 1962.

[16] Arturo Camacho and John G Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124 (3):1638–1652, 2008.

[17] Paul Taylor. Analysis and synthesis of intonation using the tilt model. *The Journal of the acoustical society of America*, 107(3):1697–1714, 2000.

[18] DSM-5 American Psychiatric Association et al. Diagnostic and statistical manual of mental disorders. *Arlington: American Psychiatric Publishing*, 2013.

[19] Sikha Bagui and Sbuhash Bagui. Jmasm17: An algorithm and code for computing exact critical values for friedman's nonparametric anova. *Journal of Modern Applied Statistical Methods*, 4(1):28, 2005.

[20] Marc Estenne, Luciano Zocchi, MICHAEL Ward, and PETER T Macklem. Chest wall motion and expiratory muscle use during phonation in normal humans. *Journal of Applied Physiology*, 68(5):2075–2082, 1990.

[21] GA Cavagna and R Margaria. An analysis of the mechanics of phonation. *Journal of Applied Physiology*, 20(2):301–307, 1965.

[22] MH Draper, Peter Ladefoged, and David Whitteridge. Expiratory pressures and air flow during speech. *British medical journal*, 1(5189):1837, 1960.

[23] Jere Mead, Arend Bouhuys, and Donald F Proctor. Mechanisms generating subglottic pressure. *Annals of the New York Academy of Sciences*, 155(1):177–182, 1968.

[24] Thomas J Hixon and Gary Weismer. Perspectives on the edinburgh study of speech breathing. *Journal of Speech, Language, and Hearing Research*, 38(1):42–60, 1995.

[25] Thomas J Hixon, Michael D Goldman, and Jere Mead. Kinematics of the chest wall during speech production: Volume displacements of the rib cage, abdomen, and lung. *Journal of Speech, Language, and Hearing Research*, 16(1):78–115, 1973.

[26] Thomas J Hixon, Jere Mead, and Michael D Goldman. Dynamics of the chest wall during speech production: Function of the thorax, rib cage, diaphragm, and abdomen. *Journal of Speech, Language, and Hearing Research*, 19(2):297–356, 1976.

[27] Richard J Morris and WS Brown. Age-related differences in speech variability among women. *Journal of Communication Disorders*, 27(1):49–64, 1994.

[28] Alison L Winkworth, Pamela J Davis, Roger D Adams, and Elizabeth Ellis. Breathing patterns during spontaneous speech. *Journal of Speech, Language, and Hearing Research*, 38(1):124–144, 1995.

[29] Doug H Whalen and Jeffrey M Kinsella-Shaw. Exploring the relationship of inspiration duration to utterance duration. *Phonetica*, 54(3-4):138–152, 1997.

[30] Alison L Winkworth, Pamela J Davis, Elizabeth Ellis, and Roger D Adams. Variability and consistency in speech breathing during readinglung volumes, speech intensity, and linguistic factors. *Journal of Speech, Language, and Hearing Research*, 37(3):535–556, 1994.

[31] Carol A Boliek, Thomas J Hixon, Peter J Watson, and Wayne J Morgan. Vocalization and breathing during the first year of life. *Journal of Voice*, 10(1):1–22, 1996.

[32] Pamela J Davis, Shi Ping Zhang, Alison Winkworth, and Richard Bandler. Neural control of vocalization: respiratory and emotional influences. *Journal of Voice*, 10 (1):23–38, 1996.

[33] David H McFarland. Respiratory markers of conversational interaction. *Journal of Speech, Language, and Hearing Research*, 44(1):128–143, 2001.

[34] Peter Ladefoged and Gerald Loeb. Preliminary studies on respiratory activity in speech. *UCLA Working Papers in Phonetics*, pages 50–60, 2002.

[35] W Tecumseh Fitch and Marc D Hauser. Unpacking "honesty": vertebrate vocal production and the evolution of acoustic signals. In *Acoustic communication*, pages 65–137. Springer, 2003.

[36] Nathalie Seddon. Ecological adaptation and species recognition drives vocal evolution in neotropical suboscine birds. *Evolution*, 59(1):200–215, 2005.

[37] David Ross Dickson and Wilma Maue-Dickson. *Anatomical and physiological bases of speech.* Little Brown and Company, 1982.

[38] Åke Randestad, Carl-Eric Lindholm, and Peter Fabian. Dimensions of the cricoid cartilage and the trachea. *The Laryngoscope*, 110(11):1957–1961, 2000.

[39] Willard R Zemlin. *Speech and Hearing Science, Anatomy and Physiology. Fourth edition.* Boston: Allyn and Bacon, 1998.

[40] Maria Schuster, Jörg Lohscheller, Peter Kummer, Ulrich Eysholdt, and Ulrich Hoppe. Laser projection in high-speed glottography for high-precision measurements of laryngeal dimensions and dynamics. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, 262(6):477–481, 2005.

[41] Minoru Hirano. Morphological structure of the vocal cord as a vibrator and its variations. *Folia Phoniatrica et Logopaedica*, 26(2):89–94, 1974.

[42] Ingo R Titze. *Principles of voice production.* Prentice Hall Englewood Cliffs, NJ:, 1994.

[43] Ira Sanders, Yingshi Han, Surinder Rai, and Hugh F Biller. Human vocalis contains distinct superior and inferior subcompartments: possible candidates for the two masses of vocal fold vibration. *Annals of Otology, Rhinology & Laryngology*, 107(10):826–833, 1998.

[44] Michael Broniatowski, David R Nelson, Robert W Shields, Sharon Grundfest-Broniatowski, Raymond Dessoffy, and Marshall Strome. Electronic analysis of intrinsic laryngeal muscles in canine sound production. *Annals of Otology, Rhinology & Laryngology*, 111(6):542–552, 2002.

[45] Janwillem Van den Berg. Myoelastic-aerodynamic theory of voice production. *Journal of speech and hearing research*, pages 227–44, 1958.

[46] Randall L Plant, Gary L Freed, and Richard E Plant. Direct measurement of onset and offset phonation threshold pressure in normal subjects. *The Journal of the Acoustical Society of America*, 116(6):3640–3646, 2004.

[47] Richard S McGowan. An aeroacoustic approach to phonation. *The Journal of the Acoustical Society of America*, 83(2):696–704, 1988.

[48] Zhaoyan Zhang. Influence of flow separation location on phonation onseta). *The Journal of the Acoustical Society of America*, 124(3):1689–1694, 2008.

[49] Sid Khosla, Shanmugam Murugappan, Randal Paniello, Jun Ying, and Ephraim Gutmark. Role of vortices in voice production: normal versus asymmetric tension. *The Laryngoscope*, 119(1):216–221, 2009.

[50] David A Berry. Mechanisms of modal and nonmodal phonation. *Journal of Phonetics*, 29(4):431–450, 2001.

[51] Wei Zhao, Cheng Zhang, Steven H Frankel, and Luc Mongeau. Computational aeroacoustics of phonation, part i: Computational methods and sound generation mechanisms. *The Journal of the Acoustical Society of America*, 112(5):2134–2146, 2002.

[52] Michael H Krane. Aeroacoustic production of low-frequency unvoiced speech sounds. *The Journal of the Acoustical Society of America*, 118(1):410–427, 2005.

[53] Sid Khosla, Shanmugam Muruguppan, Ephraim Gutmark, and Ronald Scherer. Vortical flow field during phonation in an excised canine larynx model. *Annals of Otology, Rhinology & Laryngology*, 116(3):217–228, 2007.

[54] Johan Sundberg and Thomas D Rossing. The science of singing voice. *the Journal of the Acoustical Society of America*, 87(1):462–463, 1990.

[55] David M Howard. The human singing voice. In *PROCEEDINGS-ROYAL INSTITUTION OF GREAT BRITAIN*, volume 70, pages 113–134. Oxford University Press, 1999.

[56] Gunnar Fant. Acoustic theory of speech production.'s-gravenhage: Mouton and co. *I960*, 1960.

[57] W Lawrence Gulick, George A Gescheider, and Robert D Frisina. *Hearing: Physiological acoustics, neural coding, and psychoacoustics.* Oxford University Press, 1989.

[58] David RR Smith, Roy D Patterson, Richard Turner, Hideki Kawahara, and Toshio Irino. The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America*, 117(1):305, 2005.

[59] Ronald J Baken and Robert F Orlikoff. *Clinical measurement of speech and voice.* Cengage Learning, 2000.

[60] Harry Hollien, Paul Moore, Ronald W Wendahl, and John F Michel. On the nature of vocal fry. *Journal of Speech, Language, and Hearing Research*, 9(2):245–247, 1966.

[61] Dennis M Moore and Gerald S Berke. The effect of laryngeal nerve stimulation on phonation: A glottographic study using an invivo canine model. *The Journal of the Acoustical Society of America*, 83(2):705–715, 1988.

[62] Ingo R Titze, Erich S Luschei, and Minoru Hirano. Role of the thyroarytenoid muscle in regulation of fundamental frequency. *Journal of Voice*, 3(3):213–224, 1989.

[63] Juan G Roeder. The physics and psychophysics of music, 1995.

[64] Nobuhiko Isshiki. Regulatory mechanism of voice intensity variation. *Journal of Speech, Language, and Hearing Research*, 7(1):17–29, 1964.

[65] Shinzo Tanaka and Masahiro Tanabe. Glottal adjustment for regulating vocal intensity an experimental study. *Acta oto-laryngologica*, 102(3-4):315–324, 1986.

[66] Eva B Holmberg, Robert E Hillman, and Joseph S Perkell. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *The Journal of the Acoustical Society of America*, 84(2): 511–529, 1988.

[67] Elaine T Stathopoulos and Christine Sapienza. Respiratory and laryngeal measures of children during vocal intensity variation. *The Journal of the Acoustical Society of America*, 94(5):2531–2543, 1993.

[68] Elaine T Stathopoulos and Christine Sapienza. Respiratory and laryngeal function of women and men during vocal intensity variation. *Journal of Speech, Language, and Hearing Research*, 36(1):64–75, 1993.

[69] Minoru Hirano, John Ohala, and William Vennard. The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation. *Journal of Speech, Language, and Hearing Research*, 12(3):616–628, 1969.

[70] Thomas Gay, Hajime Hirose, Marshall Strome, and Masayuki Sawashima. Electromyography of the intrinsic laryngeal muscles during phonation. *Ann Otol Rhinol Laryngol*, 81(3):401–409, 1972.

[71] Bruce R Gerratt and Jody Kreiman. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, 29(4):365–381, 2001.

[72] D Timothy Ives, David RR Smith, and Roy D Patterson. Discrimination of speaker size from syllable phrasesa). *The Journal of the Acoustical Society of America*, 118 (6):3816–3822, 2005.

[73] Kenneth N Stevens. Sources of inter-and intra-speaker variability in the acoustic properties of speech sounds. In *Proceedings of the 7th International Congress of Phonetic Sciences*, pages 1596–1607, 1971.

[74] FJD Nolan. *The phonetic bases of speaker recognition*. PhD thesis, University of Cambridge, 1980.

[75] John Laver. The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, 31:1–186, 1980.

[76] Paul L Garvin and Peter Ladefoged. Speaker identification and message identification in speech recognition. *Phonetica*, 9(4):193–199, 1963.

[77] Michael HL Hecker. *Speaker recognition: An interpretive survey of the literature*. American Speech and Hearing Assn., 1971.

[78] Douglas A Reynolds. Large population speaker identification using clean and telephone speech. *Signal Processing Letters, IEEE*, 2(3):46–48, 1995.

[79] Blas Payri. *Perception de la voix parlée: cohérence du timbre du locuteur*. PhD thesis, Universite de Paris X, 2000.

[80] Bernard Bloch and George Leonard Trager. *Outline of linguistic analysis*. Published by Linguistic Society of America at the Waverly Press, 1942.

[81] Grant Fairbanks. *Voice and Articulation: Drillbook*. Harper & Brothers, 1940.

[82] Ilse Lehiste and Gordon E Peterson. Transitions, glides, and diphthongs. *The Journal of the Acoustical Society of America*, 33(3):268–277, 1961.

[83] Peter Ladefoged and Norris P McKinney. Loudness, sound pressure, and subglottal pressure in speech. *The journal of the Acoustical Society of America*, 35(4):454–460, 1963.

[84] Björn Lindblom, Bertil Lyberg, and Karin Holmgren. *Durational patterns of Swedish phonology: do they reflect short-term motor memory processes?*, volume 3. Indiana University Linguistics Club, 1981.

[85] DONALD J SHARF. Physiologic, acoustic, and perceptual aspects of coarticulation: Implications. *Speech and language: Advances in basic research and practice*, 5:153, 1981.

[86] Paul Ekman. *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Macmillan, 2007.

[87] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[88] Klaus R Scherer and Paul Ekman. *Approaches to emotion*. Psychology Press, 2014.

[89] Klaus R Scherer. Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[90] Klaus R Scherer. Expression of emotion in voice and music. *Journal of voice*, 9 (3):235–248, 1995.

[91] Charles Darwin. *The expression of the emotions in man and animals*, volume 526. University of Chicago press, 1965.

[92] Peter Marler, KR Scherer, and P Ekman. Animal communication: affect or cognition. *Approaches to emotion*, pages 345–365, 1984.

[93] Klaus R Scherer. On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology*, 7(2):79–100, 1988.

[94] Paul Leyhausen. Biologie von ausdruck und eindruck. *Psychologische Forschung*, 31(3):177–227, 1967.

[95] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.

[96] Klaus R Scherer. Vocal affect signalling: A comparative approach. *Advances in the study of behavior*, 15:189–244, 1985.

[97] Klaus R Scherer. Vocal correlates of emotional arousal and affective disturbance. *Handbook of social psychophysiology*, pages 165–197, 1989.

[98] Dimitrios Ververidis and Constantine Kotropoulos. A state of the art review on emotional speech databases. In *Proceedings of 1st Richmedia Conference*, pages 109–119. Citeseer, 2003.

[99] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006.

[100] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[101] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[102] Marc Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, and Stan CAM Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. In *INTERSPEECH*, pages 87–90, 2001.

[103] Carl E Williams and Kenneth N Stevens. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250, 1972.

[104] A Batliner, J Buckow, H Niemann, and E Nöth. Volkerwarnke, verbmobile foundations of speech to speech translation. *ISBN 3540677836, 9783540677833: springer*, 2000.

[105] Thurid Vogt and Elisabeth André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 474–477. IEEE, 2005.

[106] Tanja Bänziger and Klaus R Scherer. The role of intonation in emotional expressions. *Speech communication*, 46(3):252–267, 2005.

[107] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011.

[108] Donna Erickson, Kenji Yoshida, Caroline Menezes, Akinori Fujino, Takemi Mochida, and Yoshiho Shibuya. Exploratory study of some acoustic and articulatory characteristics of sad speech. *Phonetica*, 63(1):1–25, 2006.

[109] Björn Schuller and Anton Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[110] Stefan Werner and Eric Keller. Prosodic aspects of speech. In *Fundamentals of speech synthesis and speech recognition*, pages 23–40. John Wiley and Sons Ltd., 1995.

[111] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1970–1973. IEEE, 1996.

[112] Chul Min Lee and Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303, 2005.

[113] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[114] Marc Schröder and Roddy Cowie. Issues in emotion-oriented computing–towards a shared understanding. In *Workshop on Emotion and Computing at KI*, 2006.

[115] Marc Schröder. Emotional speech synthesis: a review. In *INTERSPEECH*, pages 561–564, 2001.

[116] Iain R Murray and John L Arnott. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4): 369–390, 1995.

[117] Janet E Cahn. The generation of a ect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19, 1990.

[118] Iain R Murray, John L Arnott, and Elizabeth A Rohwer. Emotional stress in synthetic speech: Progress and future directions. *Speech Communication*, 20(1): 85–91, 1996.

[119] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1):227–256, 2003.

[120] Sinéad McGilloway, Roddy Cowie, Ellen Douglas-Cowie, Stan Gielen, Machiel Westerdijk, and Sybert Stroeve. Approaching automatic recognition of emotion from voice: a rough benchmark. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[121] Akemi Iida, Nick Campbell, Fumito Higuchi, and Michiaki Yasumura. A corpus-based speech synthesis system with emotion. *Speech Communication*, 40(1):161–187, 2003.

[122] Iker Luengo, Eva Navas, Inmaculada Hernáez, and Jon Sánchez. Automatic emotion recognition using prosodic parameters. In *Interspeech*, pages 493–496, 2005.

[123] Yi-hao Kao and Lin-shan Lee. Feature analysis for emotion recognition from mandarin speech considering the special characteristics of chinese language. In *InterSpeech*, 2006.

[124] Ying Wang, Shoufu Du, and Yongzhao Zhan. Adaptive and optimal classification of speech emotion recognition. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, volume 5, pages 407–411. IEEE, 2008.

[125] Shashidhar G Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K Sreenivasa Rao. Iitkgp-sesc: speech database for emotion analysis. In *Contemporary Computing*, pages 485–492. Springer, 2009.

[126] Dimitrios Ververidis, Constantine Kotropoulos, and Ioannis Pitas. Automatic emotional speech classification. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–593. IEEE, 2004.

[127] K Sreenivasa Rao, Ramu Reddy, Sudhamay Maity, and Shashidhar G Koolagudi. Characterization of emotions using the dynamics of prosodic. In *Proc. speech prosody*, volume 4. Citeseer, 2010.

[128] KS Rao, SRM Prasanna, and TV Sagar. Emotion recognition using multilevel prosodic information. In *Workshop on image and signal processing (WISP-2007)*, 2007.

[129] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63 (4):561–580, 1975.

[130] Bishnu Saroop Atal. Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, 52(6B):1687–1697, 1972.

[131] Hisashi Wakita. Residual energy of linear prediction applied to vowel and speaker recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24 (3):270–271, 1976.

[132] Alexander I Iliev and Michael S Scordilis. Spoken emotion recognition using glottal symmetry. *EURASIP Journal on Advances in Signal Processing*, 2011:2, 2011.

[133] Arun Chauhan, Shashidhar G Koolagudi, Sabin Kafley, and K Sreenivasa Rao. Emotion recognition using lp residual. In *Students' Technology Symposium (TechSym), 2010 IEEE*, pages 255–261. IEEE, 2010.

[134] Shashidhar G Koolagudi, Raghu Reddy, and K Sreenivasa Rao. Emotion recognition from speech signal using epoch parameters. In *Signal Processing and Communications (SPCOM), 2010 International Conference on*, pages 1–5. IEEE, 2010.

[135] Abdulbasit Al-Talabani, Harin Sellahewa, and Sabah Jassim. Excitation source and low level descriptor features fusion for emotion recognition using svm and ann. In *Computer Science and Electronic Engineering Conference (CEEC), 2013 5th*, pages 156–161. IEEE, 2013.

[136] P Gangamohan, Sudarsana Reddy Kadiri, Suryakanth V Gangashetty, and B Yegnanarayana. Excitation source features for discrimination of anger and happy emotions. *Tc*, 100:1, 2014.

[137] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[138] Jainath Yadav, Anshu Kumari, and K Sreenivasa Rao. Emotion recognition using lp residual at sub-segmental, segmental and supra-segmental levels. In *Communication, Information & Computing Technology (ICCICT), 2015 International Conference on*, pages 1–6. IEEE, 2015.

[139] Jacob Benesty. *Springer handbook of speech processing*. Springer Science & Business Media, 2008.

[140] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, and Wen-Yuan Liao. Combining acoustic features for improved emotion recognition in mandarin speech. In *Affective Computing and Intelligent Interaction*, pages 279–285. Springer, 2005.

[141] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng, and Charles S Chien. Feature combination for better differentiating anger from neutral in mandarin emotional speech. In *Affective Computing and Intelligent Interaction*, pages 741–742. Springer, 2007.

[142] Carl E Williams and Kenneth N Stevens. Vocal correlates of emotional states. *Speech evaluation in psychiatry*, pages 221–240, 1981.

[143] Norhaslinda Kamaruddin and Abdul Wahab. Features extraction for speech emotion. *Journal of Computational Methods in Sciences and Engineering*, 9(1, 2S1): 1–12, 2009.

[144] Daniel Neiberg, Kjell Elenius, and Kornel Laskowski. Emotion recognition in spontaneous speech using gmms. In *Interspeech*, 2006.

[145] Dmitri Bitouk, Ragini Verma, and Ani Nenkova. Class-level spectral features for emotion recognition. *Speech communication*, 52(7):613–625, 2010.

[146] Milan Sigmund. Spectral analysis of speech under stress. *IJCSNS International Journal of Computer Science and Network Security*, 7:170–172, 2007.

[147] Christer Gobl, Ailbhe Ni, et al. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1):189–212, 2003.

[148] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *INTERSPEECH*. Citeseer, 2003.

[149] Yongjin Wang and Ling Guan. An investigation of speech-based human emotion recognition. In *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*, pages 15–18. IEEE, 2004.

[150] Joy Nicholson, Kazuhiko Takahashi, and Ryohei Nakatsu. Emotion recognition in speech using neural networks. *Neural computing & applications*, 9(4):290–296, 2000.

[151] Yu Zhou, Yanqing Sun, Lin Yang, and Yonghong Yan. Applying articulatory features to speech emotion recognition. In *Research Challenges in Computer Science, 2009. ICRCCS'09. International Conference on*, pages 73–76. IEEE, 2009.

[152] Siqing Wu, Tiago H Falk, and Wai-Yip Chan. Automatic recognition of speech emotion using long-term spectro-temporal features. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–6. IEEE, 2009.

[153] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–577. IEEE, 2004.

[154] Guojun Zhou, John HL Hansen, and James F Kaiser. Nonlinear feature based classification of speech under stress. *Speech and Audio Processing, IEEE Transactions on*, 9(3):201–216, 2001.

[155] Kathleen R Merikangas, Robert Jin, Jian-Ping He, Ronald C Kessler, Sing Lee, Nancy A Sampson, Maria Carmen Viana, Laura Helena Andrade, Chiyi Hu, Elie G Karam, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Archives of general psychiatry*, 68(3):241–251, 2011.

[156] Antonio Lanata, Gaetano Valenza, Mimma Nardelli, Claudio Gentili, and Enzo Pasquale Scilingo. Complexity index from a personalized wearable monitoring system for assessing remission in mental health. *Biomedical and Health Informatics, IEEE Journal of*, 19(1):132–139, 2015.

[157] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders, text revision (DSM-IV-TR)*. American Psychiatric Association, 2000.

[158] Erin E Michalak, Lakshmi N Yatham, and Raymond W Lam. Quality of life in bipolar disorder: a review of the literature. *Health Qual Life Outcomes*, 3(1):72, 2005.

[159] Dennis A Revicki, Louis S Matza, Emuella Flood, and Andrew Lloyd. Bipolar disorder and health-related quality of life. *Pharmacoeconomics*, 23(6):583–594, 2005.

[160] Sofia Brissos. Cognitive performance and quality of life in bipolar disorder. *Canadian Journal of Psychiatry*, 53(8):517, 2008.

[161] Márcia Kauer-Sant Anna, Benício N Frey, Ana C Andreazza, Keila M Ceresér, Fernando K Gazalle, Juliana Tramontina, S Costa, Aida Santin, and Flavio Kapczinski. Anxiety comorbidity and quality of life in bipolar disorder patients. *Canadian Journal of Psychiatry*, 52(3):175, 2007.

[162] Frederick K Goodwin and Kay Redfield Jamison. *Manic-depressive illness: bipolar disorders and recurrent depression*. Oxford University Press, 2007.

[163] Alan D Lopez, Colin D Mathers, Majid Ezzati, Dean T Jamison, and Christopher JL Murray. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, 367(9524):1747–1757, 2006.

[164] Ronald C Kessler, Katherine A McGonagle, Shanyang Zhao, Christopher B Nelson, Michael Hughes, Suzann Eshleman, Hans-Ulrich Wittchen, and Kenneth S Kendler. Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: results from the national comorbidity survey. *Archives of general psychiatry*, 51(1):8–19, 1994.

[165] Leah S Kleinman, Ana Lowin, Emuella Flood, Gian Gandhi, Eric Edgell, and Dennis A Revicki. Costs of bipolar disorder. *Pharmacoeconomics*, 21(9):601–622, 2003.

[166] Hans-Ulrich Wittchen and Frank Jacobi. Size and burden of mental disorders in europe—a critical review and appraisal of 27 studies. *European neuropsychopharmacology*, 15(4):357–376, 2005.

[167] Stefano Pini, Valéria de Queiroz, Daniel Pagnin, Lukas Pezawas, Jules Angst, Giovanni B Cassano, and Hans-Ulrich Wittchen. Prevalence and burden of bipolar disorders in european countries. *European Neuropsychopharmacology*, 15(4):425–434, 2005.

[168] Yuan-Who Chen and Steven C Dilsaver. Lifetime rates of suicide attempts among subjects with bipolar and unipolar disorders relative to subjects with other axis i disorders. *Biological Psychiatry*, 39(10):896–899, 1996.

[169] Eduard Vieta, M Reinares, and AR Rosa. Staging bipolar disorder. *Neurotoxicity research*, 19(2):279–285, 2011.

[170] Darryl Daugherty, Tairi Roque-Urrea, John Urrea-Roque, Jessica Troyer, Stephen Wirkus, and Mason A Porter. Mathematical models of bipolar disorder. *Communications in Nonlinear Science and Numerical Simulation*, 14(7):2897–2908, 2009.

[171] Vikram K Yeragani, Robert Pohl, Mallika Mallavarapu, and Richard Balon. Approximate entropy of symptoms of mood: an effective technique to quantify regularity of mood. *Bipolar disorders*, 5(4):279–286, 2003.

[172] Tasha Glenn, Peter C Whybrow, Natalie Rasgon, Paul Grof, Martin Alda, Christopher Baethge, and Michael Bauer. Approximate entropy of self-reported mood prior to episodes in bipolar disorder. *Bipolar disorders*, 8(5p1):424–429, 2006.

[173] Allan Gottschalk, Mark S Bauer, and Peter C Whybrow. Evidence of chaotic mood variation in bipolar disorder. *Archives of general psychiatry*, 52(11):947–959, 1995.

[174] Allan Gottschalk, Mark S Bauer, and Peter C Whybrow. Low-dimensional chaos in bipolar disorder?—reply. *Archives of general psychiatry*, 55(3):275–276, 1998.

[175] John F Greden and Bernard J Carroll. Decrease in speech pause times with treatment of endogenous depression. *Biological Psychiatry*, 15(4):575–587, 1980.

[176] Bernard J Carroll. Speech pause time: a marker of psychomotor retardation among endogenous depressives. *Biological Psychiatry*, 16(9), 1981.

[177] Amir Muaremi, Franz Gravenhorst, Agnes Grünerbl, Bert Arnrich, and Gerhard Tröster. Assessing bipolar episodes using speech cues derived from phone calls. In *Pervasive Computing Paradigms for Mental Health*, pages 103–114. Springer, 2014.

[178] Agnes Grunerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Ohler, G Troster, Oscar Mayora, Christian Haring, and Paul Lukowicz. Smartphone-based recognition of states and state changes in bipolar disorder patients. *Biomedical and Health Informatics, IEEE Journal of*, 19(1):140–148, 2015.

[179] Zahi N Karam, Emily Mower Provost, Sushil Singh, Jennifer Montgomery, Christopher Archer, Gloria Harrington, and Melvin G Mcinnis. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4858–4862. IEEE, 2014.

[180] Franz Gravenhorst, Amir Muaremi, Jakob Bardram, Agnes Grünerbl, Oscar Mayora, Gabriel Wurzer, Mads Frost, Venet Osmani, Bert Arnrich, Paul Lukowicz, et al. Mobile phones as medical devices in mental disorder treatment: an overview. *Personal and Ubiquitous Computing*, 19(2):335–353, 2015.

[181] Andrea Guidi, Sergio Salvi, Manuel Ottaviano, Claudio Gentili, Gilles Bertschy, Danilo de Rossi, Enzo Pasquale Scilingo, and Nicola Vanello. Smartphone application for the analysis of prosodic features in running speech with a focus on bipolar disorders: System performance evaluation and case study. *Sensors*, 15(11): 28070–28087, 2015.

[182] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[183] Colin D Mathers and Dejan Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *Plos med*, 3(11):e442, 2006.

[184] J Olesen, A Gustavsson, Mikael Svensson, H-U Wittchen, and B Jönsson. The economic cost of brain disorders in europe. *European Journal of Neurology*, 19(1): 155–162, 2012.

[185] World Health Organization et al. *Preventing suicide: A global imperative*. World Health Organization, 2014.

[186] J. L. McIntosh. Usa suicide 2006: Official final data., 2006. URL http://www.suicidology.org/.

[187] Keith Hawton, Carolina Casañas i Comabella, Camilla Haw, and Kate Saunders. Risk factors for suicide in individuals with depression: a systematic review. *Journal of affective disorders*, 147(1):17–28, 2013.

[188] Jean-Pierre Lépine and Mike Briley. The increasing burden of depression. *Neuropsychiatric disease and treatment*, 7(Suppl 1):3, 2011.

[189] Thomas E Joiner Jr, Jessica S Brown, and LaRicka R Wingate. The psychology and neurobiology of suicidal behavior. *Annu. Rev. Psychol.*, 56:287–314, 2005.

[190] Alexander McGirr, Johanne Renaud, Monique Seguin, Martin Alda, Chawki Benkelfat, Alain Lesage, and Gustavo Turecki. An examination of dsm-iv depressive symptoms and risk for suicide completion in major depressive disorder: a psychological autopsy study. *Journal of affective disorders*, 97(1):203–209, 2007.

[191] J John Mann, Alan Apter, Jose Bertolote, Annette Beautrais, Dianne Currier, Ann Haas, Ulrich Hegerl, Jouko Lonnqvist, Kevin Malone, Andrej Marusic, et al. Suicide prevention strategies: a systematic review. *Jama*, 294(16):2064–2074, 2005.

[192] Keith Hawton and Kees van Heeringen. Suicide. *The Lancet*, 373(9672):1372–1381, 2015/06/13 2009. doi: 10.1016/S0140-6736(09)60372-X. URL http://dx.doi.org/10.1016/S0140-6736(09)60372-X.

[193] Max Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56, 1960.

[194] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[195] James C Mundt, Peter J Snyder, Michael S Cannizzaro, Kara Chappie, and Dayna S Geralts. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of neurolinguistics*, 20(1):50–64, 2007.

[196] Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Christine B Cha, Ronald C Kessler, and Sing Lee. Suicide and suicidal behavior. *Epidemiologic reviews*, 30 (1):133–154, 2008.

[197] Trevor Sharp and Philip J Cowen. 5-ht and depression: is the glass half-full? *Current opinion in pharmacology*, 11(1):45–51, 2011.

[198] Bernhard Luscher, Qiuying Shen, and Nadia Sahir. The gabaergic deficit hypothesis of major depressive disorder. *Molecular psychiatry*, 16(4):383–406, 2011.

[199] Michael O Poulter, Lisheng Du, Ian CG Weaver, Miklós Palkovits, Gábor Faludi, Zul Merali, Moshe Szyf, and Hymie Anisman. Gaba a receptor promoter hypermethylation in suicide brain: implications for the involvement of epigenetic processes. *Biological psychiatry*, 64(8):645–652, 2008.

[200] Yogesh Dwivedi, Hooriyah S Rizavi, Robert R Conley, Rosalinda C Roberts, Carol A Tamminga, and Ghanshyam N Pandey. Altered gene expression of brain-derived neurotrophic factor and receptor tyrosine kinase b in postmortem brain of suicide subjects. *Archives of general psychiatry*, 60(8):804–815, 2003.

[201] JM Gatt, CB Nemeroff, C Dobson-Stone, RH Paul, RA Bryant, PR Schofield, E Gordon, AH Kemp, and LM Williams. Interactions between bdnf val66met polymorphism and early life stress predict brain and arousal pathways to syndromal depression and anxiety. *Molecular psychiatry*, 14(7):681–695, 2009.

[202] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De La Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009.

[203] Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. Diagnosis of depression by behavioural signals: a multimodal approach. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 11–20. ACM, 2013.

[204] Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on MultiModal User Interfaces*, 7(3):217–228, 2013.

[205] Stefan Scherer, Giota Stratou, Mohamed Mahmoud, Jill Boberg, Jonathan Gratch, Alessandro Rizzo, and Louis-Philippe Morency. Automatic behavior descriptors for psychological disorder analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[206] James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48. ACM, 2013.

[207] Judith A Hall, Jinni A Harrigan, and Robert Rosenthal. Nonverbal behavior in clinician—patient interaction. *Applied and Preventive Psychology*, 4(1):21–37, 1996.

[208] Christina Sobin and Harold A Sackeim. Psychomotor symptoms of depression. *American Journal of Psychiatry*, 154(1):4–17, 1997.

[209] S. E. Silverman. Vocal parameters as predictors of near-term suicidal risk, 1992.

[210] Claudia Sikorski, Melanie Luppa, Hans-Helmut König, Hendrik van den Bussche, and Steffi G Riedel-Heller. Does gp training in depression care affect patient outcome?-a systematic review and meta-analysis. *BMC health services research*, 12(1):10, 2012.

[211] Alex J Mitchell, Amol Vaze, and Sanjay Rao. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, 374(9690):609–619, 2009.

[212] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. An investigation of depressed speech detection: Features and normalization. In *Interspeech*, pages 2997–3000, 2011.

[213] Alistair J Flint, Sandra E Black, Irene Campbell-Taylor, Gillian F Gailey, and Carey Levinton. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of psychiatric research*, 27 (3):309–319, 1993.

[214] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. Detection of clinical depression in adolescents' speech during family interactions. *Biomedical Engineering, IEEE Transactions on*, 58(3):574–586, 2011.

[215] Elliot Moore, Mark Clements, John W Peifer, Lydia Weisser, et al. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *Biomedical Engineering, IEEE Transactions on*, 55(1):96–107, 2008.

[216] A Nilsonne. Speech characteristics as indicators of depressive illness. *Acta Psychiatrica Scandinavica*, 77(3):253–263, 1988.

[217] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Interspeech*, pages 847–851, 2013.

[218] Andrea Carolina Trevino, Thomas Francis Quatieri, and Nicolas Malyska. Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, 2011(1):1–18, 2011.

[219] Stefan Scherer, John Pestian, and Louis-Philippe Morency. Investigating the speech characteristics of suicidal adolescents. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 709–713. IEEE, 2013.

[220] Stephen E Silverman, Marilyn K Silverman, et al. Methods and apparatus for evaluating near-term suicidal risk using vocal parameters, June 13 2006. US Patent 7,062,443.

[221] Daniel J France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and D Mitchell Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *Biomedical Engineering, IEEE Transactions on*, 47(7):829–837, 2000.

[222] Asli Ozdas, Richard G Shiavi, Stephen E Silverman, Marilyn K Silverman, and D Mitchell Wilkes. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *Biomedical Engineering, IEEE Transactions on*, 51(9):1530–1540, 2004.

[223] T Yingthawornsuk, H Kaymaz Keskinpala, D Mitchell Wilkes, Richard G Shiavi, and RM Salomon. Direct acoustic feature using iterative em algorithm and spectral energy for classifying suicidal speech. In *INTERSPEECH*, pages 766–769, 2007.

[224] Thilo Deckersbach, Darin D Dougherty, and Scott L Rauch. Functional imaging of mood and anxiety disorders. *Journal of Neuroimaging*, 16(1):1–10, 2006.

[225] Karleyton C Evans, Darin D Dougherty, Mark H Pollack, and Scott L Rauch. Using neuroimaging to predict treatment response in mood and anxiety disorders. *Annals of Clinical Psychiatry*, 18(1):33–42, 2006.

[226] Helen S Mayberg, Andres M Lozano, Valerie Voon, Heather E McNeely, David Seminowicz, Clement Hamani, Jason M Schwalb, and Sidney H Kennedy. Deep brain stimulation for treatment-resistant depression. *Neuron*, 45(5):651–660, 2005.

[227] Andrew J Niemiec and Brian J Lithgow. Alpha-band characteristics in eeg spectrum indicate reliability of frontal brain asymmetry measures in diagnosis of depression. In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pages 7517–7520. IEEE, 2006.

[228] Eric J Nestler, Michel Barrot, Ralph J DiLeone, Amelia J Eisch, Stephen J Gold, and Lisa M Monteggia. Neurobiology of depression. *Neuron*, 34(1):13–25, 2002.

[229] Timothy A Brown, Peter A Di Nardo, Cassandra L Lehman, and Laura A Campbell. Reliability of dsm-iv anxiety and mood disorders: implications for the classification of emotional disorders. *Journal of abnormal psychology*, 110(1):49, 2001.

[230] Jan H Kamphuis and Arjen Noordhof. On categorical diagnoses in dsm-v: cutting dimensions at useful points? *Psychological Assessment*, 21(3):294, 2009.

[231] V Lux and KS Kendler. Deconstructing major depression: a validation study of the dsm-iv symptomatic criteria. *Psychological medicine*, 40(10):1679–1690, 2010.

[232] María A Oquendo, Enrique Baca-García, J John Mann, and José Giner. Issues for dsm-v: Suicidal behavior as a separate diagnosis on a separate axis. 2008.

[233] Dan J Stein, Katharine A Phillips, Derek Bolton, KWM Fulford, John Z Sadler, and Kenneth S Kendler. What is a mental/psychiatric disorder? from dsm-iv to dsm-v. *Psychological medicine*, 40(11):1759–1765, 2010.

[234] David Watson. Rethinking the mood and anxiety disorders: a quantitative hierarchical model for dsm-v. *Journal of abnormal psychology*, 114(4):522, 2005.

[235] Søren Dinesen Østergaard, SOW Jensen, and Per Bech. The heterogeneity of the depressive syndrome: when numbers get serious. *Acta Psychiatrica Scandinavica*, 124(6):495–496, 2011.

[236] Martin JH Balsters, Emiel J Krahmer, Marc GJ Swerts, and Ad JJM Vingerhoets. Verbal and nonverbal correlates for depression: A review. *Current Psychiatry Reviews*, 8(3):227–234, 2012.

[237] Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, 67(3):588–597, 1996.

[238] R Michael Bagby, Andrew G Ryder, Deborah R Schuller, and Margarita B Marshall. The hamilton depression rating scale: has the gold standard become a lead weight? *American Journal of Psychiatry*, 2014.

[239] Robert D Gibbons, David C Clark, and David J Kupfer. Exactly what does the hamilton depression rating scale measure? *Journal of Psychiatric Research*, 27 (3):259–273, 1993.

[240] FC Murphy, BJ Sahakian, JS Rubinsztein, A Michael, RD Rogers, TW Robbins, and ES Paykel. Emotional bias and inhibitory control processes in mania and depression. *Psychological medicine*, 29(06):1307–1321, 1999.

[241] Gary Christopher and John MacDonald. The impact of clinical depression on working memory. *Cognitive Neuropsychiatry*, 10(5):379–399, 2005.

[242] Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(01):1–38, 1999.

[243] Jarek Krajewski, Sebastian Schnieder, David Sommer, Anton Batliner, and BjöRn Schuller. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. *Neurocomputing*, 84:65–75, 2012.

[244] Murray Alpert, Enrique R Pouget, and Raul R Silva. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1): 59–69, 2001.

[245] James C Mundt, Adam P Vogel, Douglas E Feltner, and William R Lenderking. Vocal acoustic biomarkers of depression severity and treatment response. *Biological psychiatry*, 72(7):580–587, 2012.

[246] Sylvia D Kreibig. Autonomic nervous system activity in emotion: A review. *Biological psychology*, 84(3):394–421, 2010.

[247] Paul E Croarkin, Andrea J Levinson, and Zafiris J Daskalakis. Evidence for gabaergic inhibitory deficits in major depressive disorder. *Neuroscience & Biobehavioral Reviews*, 35(3):818–825, 2011.

[248] JtK Darby and H Hollien. Vocal and speech patterns of depressive patients. *Folia Phoniatrica et Logopaedica*, 29(4):279–291, 1977.

[249] Harry Hollien. Vocal indicators of psychological stress. *Annals of the New York Academy of Sciences*, 347(1):47–72, 1980.

[250] Åsa Nilsonne and Johan Sundberg. Differences in ability of musicians and nonmusicians to judge emotional state from the fundamental frequency of voice samples. *Music Perception*, pages 507–516, 1985.

[251] Zvia Breznitz. Verbal indicators of depression. *The Journal of general psychology*, 119(4):351–363, 1992.

[252] John K Darby, Nina Simmons, and Philip A Berger. Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders*, 17(2):75–85, 1984.

[253] Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski. Automatic modelling of depressed speech: Relevant features and relevance of gender. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[254] HH Stassen et al. Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of psychiatric research*, 27(3):289–307, 1993.

[255] Åsa Nilsonne, Johan Sundberg, Sten Ternström, and Anders Askenfelt. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. *The Journal of the Acoustical Society of America*, 83(2):716–728, 1988.

[256] HH Stassen, S Kuny, and D Hell. The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology*, 8(4):303–310, 1998.

[257] F Tolkmitt, Hede Helfrich, Rainer Standke, and Klaus Rainer Scherer. Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *Journal of communication disorders*, 15(3):209–222, 1982.

[258] Michael Cannizzaro, Brian Harel, Nicole Reilly, Phillip Chappell, and Peter J Snyder. Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56(1):30–35, 2004.

[259] Thomas F Quatieri and Nicolas Malyska. Vocal-source biomarkers for depression: A link to psychomotor activity. In *Interspeech*, 2012.

[260] John D Teasdale, Sarah J Fogarty, and J Mark G Williams. Speech rate as a measure of short-term variation in depression. *British Journal of Social and Clinical Psychology*, 19(3):271–278, 1980.

[261] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. Detecting depression severity from vocal prosody. *Affective Computing, IEEE Transactions on*, 4(2):142–150, 2013.

[262] Rachelle Horwitz, Thomas F Quatieri, Brian S Helfer, Bea Yu, James R Williamson, and James Mundt. On the relative importance of vocal source, system, and prosody in human depression. In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[263] Heiner Ellgring and Klaus R Scherer. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20(2):83–110, 1996.

[264] Roman Kotov, Wakiza Gamez, Frank Schmidt, and David Watson. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological bulletin*, 136(5):768, 2010.

[265] HH Stassen, G Bomben, and E Günther. Speech characteristics in depression. *Psychopathology*, 24(2):88–105, 1991.

[266] Hamish P Godfrey and Robert G Knight. The validity of actometer and speech activity measures in the assessment of depressed patients. *The British Journal of Psychiatry*, 145(2):159–163, 1984.

[267] Patrick Hardy, Roland Jouvent, and Daniel Widlöcher. Speech pause time and the retardation rating scale for depression (erd): Towards a reciprocal validation. *Journal of Affective Disorders*, 1984.

[268] E Szabadi, CM Bradshaw, and JA Besson. Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression. *The British Journal of Psychiatry*, 129(6):592–597, 1976.

[269] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, Gordon Parker, et al. From joyous to clinically depressed: Mood detection using spontaneous speech. In *FLAIRS Conference*, 2012.

[270] Åsa Nilsonne. Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatrica Scandinavica*, 76(3):235–245, 1987.

[271] Michael B First, Robert L Spitzer, Miriam Gibbon, and Janet BW Williams. *Structured Clinical Interview for DSM-IV Axis I Disorders: Patient Edition (February 1996 Final), SCID-I/P*. Biometrics Research Department, New York State Psychiatric Institute, 1998.

[272] Matti Airas. Tkk aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):49–64, 2008.

[273] Boris Doval, Christophe d'Alessandro, and Nathalie Henrich. The spectrum of glottal flow models. *Acta acustica united with acustica*, 92(6):1026–1046, 2006.

[274] Jacqueline Walker and Peter Murphy. A review of glottal waveform analysis. In *Progress in nonlinear speech processing*, pages 1–21. Springer, 2007.

[275] Anita McAllister, Elisabeth Sederholm, Sten Ternström, and Johan Sundberg. Perturbation and hoarseness: a pilot study of six children's voices. *Journal of Voice*, 10(3):252–261, 1996.

[276] Robert F Orlikoff and Joel C Kahane. Influence of mean sound pressure level on jitter and shimmer measures. *Journal of voice*, 5(2):113–119, 1991.

[277] John Laver, Steven Hiller, and Janet Mackenzie Beck. Acoustic waveform perturbations and voice disorders. *Journal of Voice*, 6(2):115–126, 1992.

[278] John Kane, Irena Yanushevskaya, John Dalton, Christer Gobl, and Ailbhe Ní Chasaide. Using phonetic feature extraction to determine optimal speech regions for maximising the effectiveness of glottal source analysis. In *INTERSPEECH*, pages 29–33, 2013.

[279] John Kane, Matthew Aylett, Irena Yanushevskaya, and Christer Gobl. Phonetic feature extraction for context-sensitive glottal source processing. *Speech Communication*, 59:10–21, 2014.

[280] Stefan Scherer, Giota Stratou, and Louis-Philippe Morency. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 135–140. ACM, 2013.

[281] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2):109–118, 1992.

[282] John HL Hansen. Analysis and compensation of stressed and noisy speech with application to robust automatic recognition. *Signal Processing*, 17(3):282, 1989.

[283] Michael HL Hecker, Kenneth N Stevens, Gottfried von Bismarck, and Carl E Williams. Manifestations of task-induced stress in the acoustic speech signal. *The Journal of the Acoustical Society of America*, 44(4):993–1001, 1968.

[284] Nicholas Cummins, Julien Epps, and Eliathamby Ambikairajah. Spectro-temporal analysis of speech affected by depression and psychomotor retardation. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7542–7546. IEEE, 2013.

[285] Brian S Helfer, Thomas F Quatieri, James R Williamson, Daryush D Mehta, Rachelle Horwitz, and Bea Yu. Classification of depression state based on articulatory precision. In *Interspeech*, pages 2172–2176. Citeseer, 2013.

[286] T Yingthawornsuk, H Kaymaz Keskinpala, D France, D Mitchell Wilkes, Richard G Shiavi, and RM Salomon. Objective estimation of suicidal risk using vocal output characteristics. In *Ninth International Conference on Spoken Language Processing*, 2006.

[287] Anne-Maria Laukkanen, Eva Björkner, and Johan Sundberg. Throaty voice quality: subglottal pressure, voice source, and formant characteristics. *Journal of Voice*, 20(1):25–37, 2006.

[288] Nicholas Cummins, Julien Epps, Vidhyasaharan Sethu, Michael Breakspear, and Roland Goecke. Modeling spectral variability for the classification of depressed speech. In *Interspeech*, pages 857–861, 2013.

[289] Douglas E Sturim, Pedro A Torres-Carrasquillo, Thomas F Quatieri, Nicolas Malyska, and Alan McCree. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Interspeech*, pages 2981–2984, 2011.

[290] Kuan Ee Brian Ooi, Margaret Lech, and Nicholas B Allen. Multichannel weighted speech classification system for prediction of major depression in adolescents. *Biomedical Engineering, IEEE Transactions on*, 60(2):497–506, 2013.

[291] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Tom Gedeon, Michael Breakspear, and Gordon Parker. A comparative study of different classifiers for detecting depression from spontaneous speech. In *Acoustics, Speech*

and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8022–8026. IEEE, 2013.

[292] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[293] J. Kominek and A. Black. Cmu arctic databases for speech synthesis cmu language technologies institute. *Language Technologies Institute, CMU, Pittsburgh PA, Tech Report CMU-LTI-03-177*, 2003.

[294] R Paradiso, AM Bianchi, K Lau, and EP Scilingo. Psyche: personalised monitoring systems for care in mental health. In *Engineering in Medicine and Biology Society (EMBC), 2010 annual international conference of the IEEE*, pages 3602–3605. IEEE, 2010.

[295] H.A. Murray. Uses of the thematic apperception test. *The American journal of psychiatry*, 107(8):577–581, 1951.

[296] Dennis H Klatt and Laura C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[297] Ken Steiglitz. *A digital signal processing primer, with applications to digital audio and computer music*. Addison Wesley Longman Publishing Co., Inc., 1997.

[298] Sascha Disch and Udo Zölzer. Modulation and delay line based digital audio effects. In *2nd Workshop on Digital Audio Effects DAFx*. Citeseer, 1999.

[299] Udo Zölzer. *Digital audio signal processing*. John Wiley & Sons, 2008.

[300] N. Vanello, N. Martini, M. Milanesi, H. Keiser, M. Calisti, L. Bocchi, C. Manfredi, and L. Landini. Evaluation of a pitch estimation algorithm for speech emotion recognition. In *Proc. 6th int. workshop models and analysis of vocal emissions for biomedical applications*, pages 29–32, 2009.

[301] Bishnu S Atal and Lawrence R Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(3):201–212, 1976.

[302] José-Luis Blanco, Jean Schoentgen, and Claudia Manfredi. Vocal tract settings in speakers with obstructive sleep apnea syndrome. In *Proc. 8th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 211–214. Firenze University Press, 2013.

[303] Nivja H De Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390, 2009.

[304] H Lazarus. Prediction of verbal communication in noise—a development of generalized sil curves and the quality of communication (part 2). *Applied Acoustics*, 20(4):245–261, 1987.

[305] Murtaza Bulut and Shrikanth Narayanan. On the robustness of overall f0-only modifications to the perception of emotions in speech. *The Journal of the Acoustical Society of America*, 123(6):4547–4558, 2008.

[306] Juan Pablo Arias, Carlos Busso, and Nestor Becerra Yoma. Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech & Language*, 28(1):278–294, 2014.

[307] K Sreenivasa Rao, Shashidhar G Koolagudi, and Ramu Reddy Vempada. Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, 16(2):143–160, 2013.

[308] Haibo Wang, Aijun Li, and Qiang Fang. F0 contour of prosodic word in happy speech of mandarin. In *Affective Computing and Intelligent Interaction*, pages 433–440. Springer, 2005.

[309] Maocan Lin. On production and perception of boundary tone in chinese intonation. In *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 2004.

[310] J Sundberg, Perkins WH Ternström, and P Gramming. Long-term-average spectrum analysis of phonatory effects of noise and filtered auditory feedback, 1987.

[311] Anne-Maria Laukkanen, Erkki Vilkman, Paavo Alku, and Hanna Oksanen. On the perception of emotions in speech: the role of voice quality. *Logopedics Phonatrics Vocology*, 22(4):157–168, 1997.

[312] Paul Boersma and Gordana Kovacic. Spectral characteristics of three styles of croatian folk singing. *The Journal of the Acoustical Society of America*, 119(3): 1805–1816, 2006.

[313] Patrick A Naylor, Anastasis Kounoudes, Jon Gudnason, and Mike Brookes. Estimation of glottal closure instants in voiced speech using the dypsa algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):34–43, 2007.

[314] Sean A Fulop. *Speech spectrum analysis*. Springer, 2011.

[315] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[316] Keelan Evanini and Catherine Lai. The importance of optimal parameter setting for pitch extraction. *The Journal of the Acoustical Society of America*, 128(4): 2291–2291, 2010.

[317] John D Markel. The sift algorithm for fundamental frequency estimation. *Audio and Electroacoustics, IEEE Transactions on*, 20(5):367–377, 1972.

[318] Maciej Pakosz. Attitudinal judgments in intonation: Some evidence for a theory. *Journal of Psycholinguistic Research*, 12(3):311–326, 1983.

[319] Rainer Banse and Klaus R Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614, 1996.

[320] Eva Gilboa-Schechtman, Lior Galili, Yair Sahar, and Ofer Amir. Being "in" or "out" of the game: subjective and acoustic reactions to exclusion and popularity in social anxiety. *Frontiers in human neuroscience*, 8, 2014.

[321] Giulio Perugi, Pierpaolo Medda, João Reis, Salvatore Rizzato, Michela Giorgi Mariani, and Mauro Mauri. Clinical subtypes of severe bipolar mixed states. *Journal of affective disorders*, 151(3):1076–1082, 2013.

[322] Michael E Thase and Thomas L Schwartz. Choosing medications for treatment-resistant depression based on mechanism of action. *The Journal of clinical psychiatry*, 76(6):720–727, 2015.

[323] Lauren WM Swager and Susan K Morgan. Psychotropic-induced dry mouth: Don't overlook this potentially serious side effect. *Current Psychiatry*, 10(12):54, 2011.

[324] Thomas F Cleveland, Johan Sundberg, and RE Stone. Long-term-average spectrum characteristics of country singers during speaking and singing. *Journal of Voice*, 15(1):54–60, 2001.