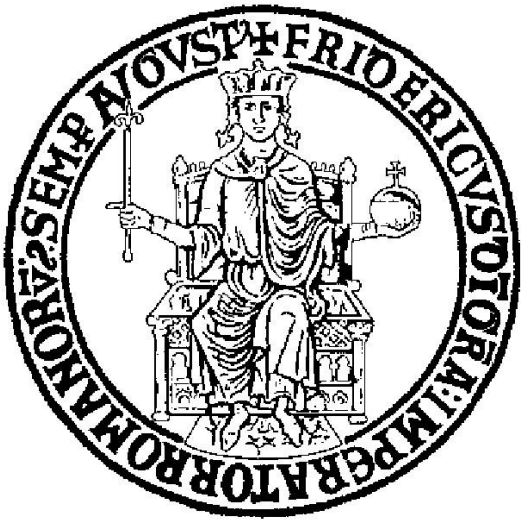
Syllable based speech analysis

for affective robotics



Antonio Origlia

University of Naples “Federico II”

Dissertation submitted for the degree of

Doctor of Philosophy

April 2013

It's your fiction that interests me. Your studies of the interplay of human motives and emotion.

ISAAC ASIMOV

I, Robot

Contents

Introduction

Curiosity is a powerful force pushing human beings towards exploring and experimenting and it is indeed the best weapon we have been using to win our place against other species in the evolutionary pit. The role of curiosity has been hailed since ancient times by Homer in his epic poems in which Odysseus' sharpness of mind, capable of winning wars brute force could not resolve, is strongly linked with his continuous challenge to the unknown. While our minds grew more and more sophisticated, however, curiosity about things and places has been integrated by curiosity about ourselves. As narcisistic as it can sound, we are often awed by our own nature. As disturbing as it can sound, on the other hand, the most scary things we meet during our lives come from the deeps of our own souls.

When it comes to understanding human nature, we often confront ourselves with well known experiences that we strive to define: we are put in charge of a machine, our body, we do not fully understand. A machine that, sometimes, behaves differently from we expected or even from what we wanted. The main primordial energy fueling this machine are indeed emotions. While playing a critical role for survival of the species and accounting for a number of situations needing fast response, emotions are pure instinct and, by definition, irrational. When the emotional self takes over the rational self, we are spoiled of any control over our actions but not of our responsability of them. Acting emotionally is tempting as it often gives satisfaction in the short term but always has consequences in the long term. Shogun Tokugawa Ieyasu, who united Japan under his rule after fighting many years against other japanese clans, knew very well how emotions can be harmful to long and delicate tasks like strategic warfare. His definition of patience, in particular, was

tightly linked to the concept of emotions:

> "The strong manly ones in life are those who understand the meaning of the word patience. Patience means restraining one's inclinations. There are seven emotions: joy, anger, anxiety, adoration, grief, fear, and hate, and if a man does not give way to these he can be called patient."

Other than a first example of discrete representation of emotions, which I will discuss in Chapter 1, it is interesting to see how emotions were considered by a great Japanese strategist to be opposing the patience needed to pursue his plan. It is right, in my opinion, to give in to emotional behavior from time to time as it is part of our nature but it is also necessary to watch over our emotional self as much is the damage it is able to bring to ourselves and to others. By residing deep inside our brain, emotions are often taken for granted. However, although we all know what we are talking about when we discuss emotions, their nature is elusive and difficult to describe, thus making our own emotions mysterious and almost unpredictable. H. P. Lovecraft used to call *unspeakable horrors* things one could not only understand but even name. This uncapability to confront with something, in a lovecraftian setup, drives people to madness. How should we feel, then, considering that the least understood thing on earth, ourselves, is constantly following us?

Questions thousands of years old about the *self* have been asked and we tried to answer them by means of the best instruments we had available throughout our relatively short presence on the planet: philosophy, religion, biology and psychology among these. We now live in the age of technology and, through this work, I make my own journey among the mists of my mind, trying to understand more about myself by using the instruments given to us by modern age knowing that, especially in this case, the journey is more important than the destination.

This work is about both human and synthetic emotions and is motivated by my opinion that it may be possible to better understand emotions by trying to simulate them. Emotions, however, do not exist in one's body only. Perceived emotions are often more important than felt emotions as they are, more or less consciously, transmitted during

communication. Here, I designed my approach by considering one of the most basic human communication channels, voice, and limiting my analysis to the acoustic properties of the human voice, without introducing semantics. This is because my attempt is aimed at understanding how much information can be gained from the raw sound without involving higher cognitive layers. I employ a set of basic linguistic rules in order to create an intepretable representation of speech that I use to discuss the results I present. This is because I will try to take advantage of the terminology coming from decades of linguistic studies on intonation to better illustrate my goals and the results I obtained. This representation, described in Chapter 2, has also an impact on the performance of the technological artifact produced as an application example of the analysis method. Since the representation I will describe identifies only specific areas of the speech signal to be important for emotion recognition and since these areas are not difficult to isolate, the required comuputation load is reduced by extracting information from these segments only. After presenting the analysis method, in Chapter 3 I will present experiments to evaluate its performance on dimensional and continuous emotional tracking by using emotional speech corpora and human annotations. However, while I believe that emotional speech corpora are useful to study emotion expression through voice, I also believe that, being hard for humans to talk about and to quantify the experienced emotional response, it may be better to evaluate the capability of a technological artifact to recognize, simulate and exploit emotions by setting up a basic interactive task in which successful emotional communication is essential to reach the designated goal. In Chapter 4, I will therefore present how the offline analysis method presented in Chapter 2 and the results presented in Chapter 3 can be ported in a real-time setting by keeping the linguistically inspired setup. The real-time implementation of the speech analysis method is used to design a simple task to check if the user's intended emotions can be correctly captured and interpreted to produce believable emotional behaviours in an animal-like robot. Coherence in the behavior of the robot can then be interpreted as good emotional recognition performance.

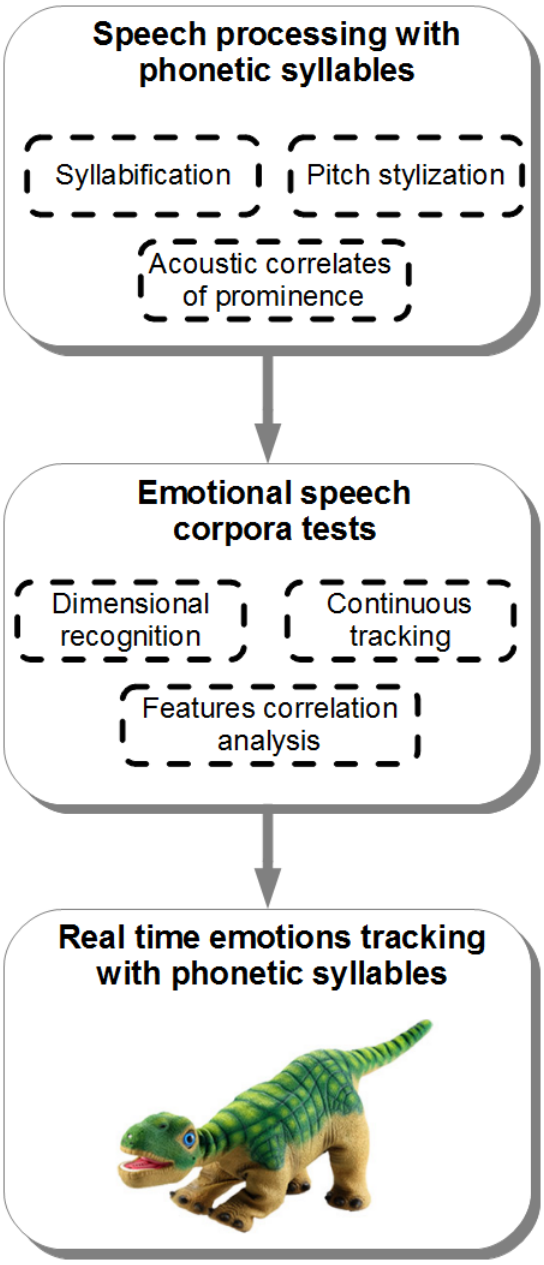


Figure 1: General organization of the presented work

Chapter 1

Emotions and Affective Computing

In this Chapter I discuss about the structure of the brain and about its reactions to external stimuli. The goal of this overview is to clarify how the brain is organized from a biological point of view, how it processes signals coming from the peripheral nervous system and how it stores past experience. Special attention will be paid to the way physiological reactions to emotional experience spread throughout the brain making emotions an important driving force with respect to cognitive processes.

1.1 Neurobiology of emotions

One of the most influential works about emotions from a neurobiological point of view is the work of Ledoux (1998). In his book, Ledoux presents a summary of his experience in studying emotions in the brain, their role in everyday life and their relationship with other mental processes. Among the themes discussed by the author in his book, I will build my approach on a subset of them.

First of all, Ledoux highlights how the brain does not possess nor does anything that can be referred to as *emotion*. This term is rather an exemplification label, a linguistic trick to enable us to talk about complex physical experiences. Being these experiences composed of many different physical reactions like faster heartbeat, increased breathing

frequency and modifications in muscular tension, it makes sense to assume that there are different systems, each one dedicated to the control of an emotional reaction, that work together to create the experience we refer to as *emotion*.

Another point of interest for Ledoux is that emotions cannot be controlled consciously. We can control external elements in such a way that we know a certain emotion will arise but we can't simply experience that emotion. While emotions cannot be rationally controlled, they influence rational thought. They consist of basic contextual evaluations that can both support decision processes or even establish goals for them if we follow the theory presented in Arnold (1960) that emotions are the tendency to go towards what we believe to be useful and to stay away from what we believe it is harmful.

Ledoux also highlights a number of neurobiologically motivated reasons why emotions cannot be assimilated to cognition but should rather be considered a different system *interacting* with cognitive processes:

- Damages to specific regions of the brain can erase the capability of emotionally evaluating a stimulus without altering the capability of perceiving the stimulus. Representation and evaluation are therefore processes that are performed separately by the brain.

- Emotional evaluation can be completed before cognitive processing is complete. In other terms, emotional significance can be attributed to an object before knowing what it is.

- Emotional memory is stored separately from cognitive memory. Specific kinds of damage can remove the capability of assigning emotional value to a stimulus without altering the capability of remembering information about previous experiences with that stimulus and vice-versa.

- Emotional evaluation systems are directly connected to action systems. Cognitive processing systems are not so tightly bound to reaction systems.

- Being strongly associated with reaction systems, it is more frequent that results of an emotional evaluation process provoke physical sensations than what happens with cognitive processes.

To support the distinction between emotion and cognition, I would like to add to the observations made by Ledoux a personal one. It is common, in nature, to observe defence strategies based on mimicry stimuli usually associated with dangerous species. This is called *Batesian mimicry*. For example, a cat's hissing has been described as an imitation of the snake's menacing sound. As reptilians are among the most ancient creatures on Earth, it makes sense that their warnings are so very well known by all other terrestrial creatures to be automatically associated with a *danger* feeling. For an relatively recent creature like a mammal, it makes sense to use a hiss as a warning or a menace by exploiting the instinctive *fear* response associated to that sound. This works in practice although a cat is clearly not a snake. Other species, moreover, evolved to disguise themselves as more dangerous creatures. The *anilius scytale* is a harmless snake that disguises itself by taking the same colors, but with a different pattern, of the venomous *Micrurus fulvius*, better known as coral snake. For this reason, the *anilius scytale* is also known as the false coral snake. In this case, the *fear* reaction associated with the colors of the coral snake are exploited by the false coral snake to survive. Through cognitive processes we are, of course, able to distinguish the two species because of the different pattern but the fast, emotional, *fear* reaction we experience regardless of our reasoning is what the snake relies onto in order to be able to scare off potential predators and flee. This adds, in my opinion, on LeDoux observations regarding the separation of emotional evaluation systems from cognitive processes I will follow in this work.

1.2 Psychology of emotions

Different models have been proposed in psychology to describe emotions in a formal way. In this Section I will present the most important models that have been proposed in the

Figure 1.1: A subset of the images used in Darwins investigation

literature: categorical, dimensional and appraisal models.

1.2.1 Categorical models

Deriving directly from the usual way we talk about emotions by labeling them, emotions can be separated into discrete classes. The first discrete representation of emotions was formulated by Charles Darwin (Darwin, 1872). In his dissertation about genetically determined communication concerning facial expression, Darwin dedicates a book chapter to each of the superclasses he identified among emotions: low spirits like grief, high spirits like love, then anger, disgust, surprise and, lastly, complex emotions like shame and shyness. Darwin relied on a questionnaire he circulated among different ethnic groups along with a large number of pictures showing emotionally expressive actors and children and on descriptions of psychiatric patients. His work aimed at identifying the common traits between humans and animals in the expression of basic feelings and emotions. Figure 1.1 shows a sample of the images used by Darwin for his investigation on emotions.

In Ekman (1992), a study specifically aimed at identifying a set of basic emotions to which all other emotions could be referred to was presented. As in Darwin's work, the

Figure 1.2: Facial expressions associated with Ekman's basic emotions

experiment was based on facial expressions evaluated by different ethnic groups and it identified a set of six emotions that which expression was shared among cultures. These were anger, joy, sadness, disgust, fear and surprise (the neutral emotion was also considered). Figure 1.2 shows the set of facial expressions associated with Ekman's basic emotions.

1.2.2 The dimensional model

Research on emotions has moved from discrete classifications through categories to a more dynamical representation using multi-dimensional spaces, where the focus is on the components of emotions rather than on emotions themselves. At the very first stage in emotion research, the choice for discrete (and therefore basic and acted) emotions, considered as categories, was mainly due to the fact that such material was easier to obtain and to collect (Schuller et al., 2011). On the other hand, the increasing development and interest in human-machine interactions with more realistic and real-time situations forced researchers to take into account real-life variations in landmark emotions by using dynamical scales (dimensions) in order to introduce more variability in the classification of emotional speech. For this reason researcher advocating the dimensional description of emotions argue that a single label or a set of discrete classes may not be able to account for the complexity and the variability of affective information. According to the dimensional view of human

affect, affective states are not independent from one another but they are related to one another in a systematic manner (Gunes et al., 2011).

Dimensional descriptions of emotions usually adopt annotation scales along a continuum of affective behavior in terms of latent dimensions (eg. arousal, power and valence). A number of different models accounting for a dimensional description of human affect exist. One of these models, the *circumplex model of affect* introduced by Russell (1980), represents emotions as a bipolar entity of arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant), therefore being part of the same emotional continuum. Cowie et al. (2000, 2001) used a similar model to model and assess affect from speech. The 3D emotional space proposed by Mehrabian (1996) is another dimensional model accounting for pleasure - displeasure, arousal - non arousal and dominance - submissiveness (also referred to as PAD or as *emotional primitives*; see Espinosa et al. (2010); Jia et al. (2011). Grimm and Kroschel (2005) used a similar framework to describe emotions by means of three emotion *primitives*, or attributes (valence, activation, and dominance) proposing a real-valued 3D emotion space concept to overcome the limitations of discrete emotion categorization (also referred to as VAD space). However, as noted by (Gunes et al., 2011, p.829)

> "[. . .] there is no coding scheme that is agreed upon and used by all researchers in the field that can accommodate all possible communicative cues and modalities."

This absence of agreement on a coding scheme in the dimensional view of emotions and the variability of scales and attributes considered falls back into the availability of speech databases. As far as naturalistic databases are concerned (as opposed to acted) the most employed database is The Vera am Mittag (VAM) corpus collected by Grimm et al. (2008) describing emotions on a continuous-valued scale in the valence, activation, and dominance (VAD) space.

From a theoretical point of view, Fontaine et al. (2007) emphasize the considerable disagreement on how many dimensions are essential to provide an optimal framework in emotion research and point out how the debate still remains open. However, the research

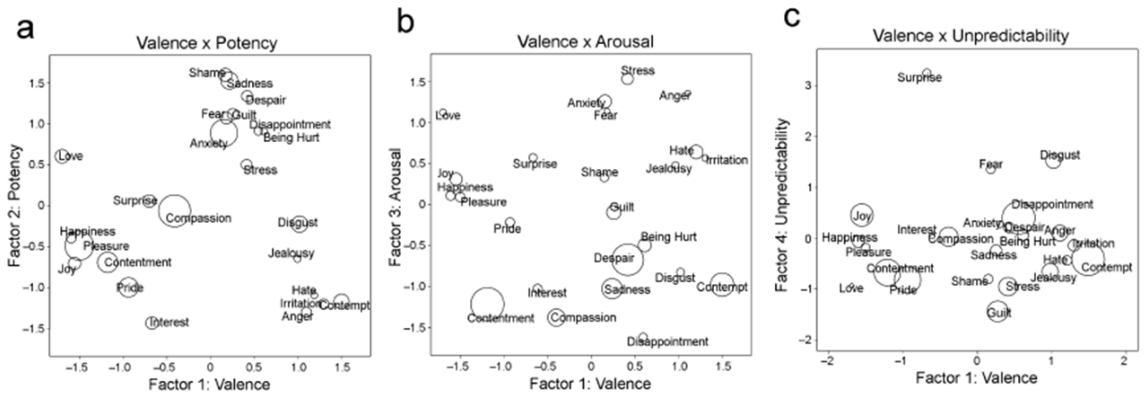


Figure 1.3: Emotional labels distributed in the four dimensions indicated by Fontaine et al. (2007)

carried out by Fontaine et al. (2007) in the representation of the semantic space of emotion confirms that more than two dimensions are needed for a low-dimensional representation. In proposing a four-dimensional structure they also confirm that the three dimensions (evaluation-pleasantness, potency-control, activation-arousal) taken into account in the first studies in this domain are the most important ones even in a multi-cultural setting. This interest in the way of representing emotions is not limited to the theoretical field but it is a very important topic from the technical point of view too (Wöllmer et al., 2008). Despite the debate regarding the opportunity of using a discrete rather than a dimensional representation of emotions, the two directions are not mutually exclusive and, as it was noted in (Schuller et al., 2011, p.1065)

> "[...] Irrespective of strong beliefs in the one or the other type of modeling, in practice, categories can always be mapped onto dimensions and vice versa albeit not necessarily lossless."

In Figure 1.3, the relation between emotional labels and the model presented in Fontaine et al. (2007) is shown.

While a categorical view of emotions easily matches the discrete nature of machines, a dimensional model allows designers to decompose the complex phenomenon represented by emotions into its components. Later discretization can be performed in the n-dimensional space if needed.

1.2.3 Appraisal models

Appraisal models are based on the concept that emotions arise from the evaluation of contingent events in terms of present and future effects both concerning the subject and other involved people. A number of appraisal models of emotions have been proposed in the literature. Due to the difficulty of defining appraisal in general terms, these can be significantly different as noted in (Scherer et al., 2001, p. 11)

> "Examination of these models indicates that although there is significant overlap [between the two types of structural models], there are also differences: in which appraisals are included; how particular appraisals are operationalized; which emotions are encompassed by a model; and which particular combinations of appraisals are proposed to elicit a particular emotional response."

Among these approaches, Roseman's theory of appraisal (Roseman, 1996) and Scherer's Multi Level Sequential Checks (Scherer et al., 2001). Roseman's appraisal theory describes emotions as the composition of the outcomes of specific evaluations related to the experienced situation and considers motive consistency and accountability as the two most important components of the appraisal process. Figure 1.4 shows a summary of the emotions considered by Roseman and of their composition relatively to the relevant evaluations. Sadness, for example, is elicited by verified circumstances aversing the subject's goal and on which control potential is low

Multi Level Sequential Checks are made up of three levels of appraisal process, with sequential constraints at each level of processing that create a specifically ordered processing construct happening ad different conscience levels. The first of these levels describes mechanisms that are mostly genetically determined like prototypic unconditioned fear eliciting stimuli. The second level describes socially learned behaviors, which become almost automatic like in the first level but are not innate. The last level refers to emotions elicitated by high level, propositional-symbolic processes related to goals and beliefs. These levels of appraisal and their components are summarized in Table 1.1 as they were reported in

	Unexpected-ness	Motivational State and Situational State		Probability	Control Potential	Agency	Problem type
		Motivation	Motive-Consistency				
Surprise	Unexpected	-	-	-	-	-	-
Joy	-	Appetitive	Consistent	Certain	-	Circumstance-Caused	-
Relief	-				-		-
Sadness	-	Aversive	Inconsistent		Low control potential		-
Distress	-						-
Hope	-	-	Consistent	Uncertain	-		-
Fear	-	-	Inconsistent		Low control potential		-
Liking	-	-	Consistent	-	-	Other-Caused	-
Dislike	-	-	Inconsistent	-	Low control potential		-
Pride	-	-	Consistent	-	-	Self-Caused	-
Regret	-	-	Inconsistent	-	Low control potential		-
Frustration	-	-		-	High control potential	Circumstance-Caused	Non-Characterological
Anger	-	-		-		Other-Caused	
Guilt	-	-		-		Self-Caused	
Disgust	-	-		-		Circumstance-Caused	Characterological
Contempt	-	-		-		Other-Caused	
Shame	-	-		-		Self-Caused	

Figure 1.4: Roseman's appraisal schema

Level	Novelty	Pleasantness	Goal/need conductiveness	Coping potential	Norm/self compatibility
Sensory-motor	Sudden, intense stimulation	Innate preferences / aversions	Basic needs	Available energy	Empathic adaptation (?)
Schematic	Familiarity (schema matching)	Learned preferences / aversions	Acquired needs / motives	Body schema	Self / social schemata
Conceptual	Expectations (cause / effect, probability)	Anticipated positive / negative estimates	Conscious goals	Problem solving ability	Self ideal, moral evaluation

Table 1.1: Appraisal levels as presented by Leventhal and Scherer (1987)

(Leventhal and Scherer, 1987, p. 17), which is the work Scherer et al. (2001) builds on.

Appraisal models represent a tendency to treat emotions like other cognitive processes. Ledoux (1998), on the basis of the neurobiologically motivated points presented in Section 1.1, is against the *forceful* inclusion of emotions in a cognitivist setup by stating (Ledoux, 1998, p. 68-69):

> "My desire to protect emotion from being consumed by the cognitive monster comes from my understanding of how emotion is organized in the brain."

This particularly strong expression is motivated by the attempt of cognitivists to include emotions in a *cold* framework in which logical processes are used to describe even the most innate reactions. Marvin Minsky, pioneering leader of this research field stated that emotions are (Minsky, 2006, p. 1)

> "[. . .] not especially different from the processes that we call thinking."

In his dissertation, Ledoux acknowledges that appraisal theories are very close to the truth but he criticizes two points of the approach. First of all, he considers the investigation method, which is mainly based on verbally reported introspection by subjects, to be inappropriate to study emotions as they are usually treated, in language, through exemplification labels taht do not contain all the details of *what* an emotion actually is and *why* it arises. Secondly, he considers cognitive processes to have too much weight in emotion definition.

For what it concerns the influence of cognitive processes, I personally agree with Ledoux that a strictly cognitivist view of emotions is to be avoided. On the other hand, I believe that much of the debate is caused by the unclear use of emotional terms and by the absence of a prudent distinction, from a cognitive point of view, of what is intended to be described through cognitive processes and what actually *is* a cognitive process.

To discuss my views on this matter, I will concentrate on Scherer's Multi Level Sequential Checks theory. Concerning the first point, I believe that it is misleading to talk about emotions on all the three levels considered in the model. I believe that different terms should be used to describe the results of the evaluation performed on each level. The results of higher level evaluation, involving goals, beliefs and personal experiences produce *feelings* and, being this level strictly associated with symbolic computation, it is the level in which a strictly cognitivist view is appropriate.

Results obtained through second level evaluations, being determined by social experience, produce *affections*. I choose this term by following Spinoza's concept of *affectus* as the modification or variation produced in a body (including the mind) when it interacts with another body *which increases or diminishes the body's power of activity*.

On the first level there is what I think should be called *emotions* as the purely reactive nature of the outcomes of this kind of processing appears to me to be particularly close to the meaning of commonly used emotional labels.

The Multi Level Sequential Checks theory accounts for instinctive reactions on this first level but it includes an explicit evaluation of events against *basic needs* and *innate preferences* through cognitive processes. As these processes are said to be genetically determined and automatic, the presence of a motivation in the reaction appears to be out of place. This, however, is more an artifact than a real problem and it is caused by the need of presenting a uniform model: although we are not capable of verbally explaining the reasons why we experience an emotional reaction, this does not mean that there is no motivation. The reason why we have an innate fear snakes, for example, is motivated by the evolutionary need of protecting ourselves from venomous species. While, at present, we do not need to think about the possible danger represented by a snake to experience fear,

the process associating the general template of the snake or the sounds it produces can be described in cognitive terms but it really represents an evolutionary process. That is, I believe that as long as cognitivism does not claim to represent what *actually happens* in the brain at lower levels and limits itself to describe an interiorized, evolutionarily determined process in cognitivist terms with the only goal of pursuing model uniformity, there is enough room for the different views to coexist. Describing emotions in cognitive terms is acceptable in my opinion, as long as the motivations coming with cognitive processing are attributed to evolution rather than to conscience.

In the view of appraisal theories, I will concentrate here of the first level only, considering paralinguistic features for emotions communication and a reactive robotics architecture to evaluate the coherence of the behaviors shown on the basis of automatic recognition.

1.3 Affective computing

The idea of including emotional behavior into machines has been proposed since the beginning of research into artificial intelligence. Affective computing is a branch of artificial intelligence that has gained significant importance, in recent years, mainly because of the work by Rosalind Picard (Picard, 1997), who defined the term too. A formal definition of the term is reported in Tao and Tieniu (2005):

> Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects.

Emotions have been found to be critical in order to obtain believable artificial agents both in the form of conversational systems and in robotics. Of course, the field includes many different areas of application because of its broad definition. In this Section, I will concentrate on summarizing the work that has been done in recent years for what it concerns emotional speech (excluding semantic analysis) and robotics.

1.3.1 Emotional speech

Studying the way emotions are conveyed through speech is probably the main field of application of affective computing together with facial expression recognition. Multimodal databases of emotionally colored content mainly account for these two modalities with gestures gaining more importance in recent years. Studying emotions is complicated mainly because of the need researchers have of big amounts of descriptive data. Being emotions hard to describe, however, makes it difficult to collect this kind of material and to represent its content. Categorically annotated corpora have been used in the first years of computationally based analysis of emotional speech (e. g. the Berlin Emotional Speech Database (Burkhardt et al., 2005)) while, in recent years, dimensional models have been used instead (e. g. the Vera Am Mittag corpus (Grimm et al., 2008)). Other than annotation, elicitation methods have been discussed. Read or acted speech, used initially, has been found to be significantly different from spontaneous speech (Vogt and Andre, 2005; Jürgens et al., 2011). While acted speech has been useful in the first years to study the basics of emotion communication through speech, research has now consistently moved to spontaneous emotions. These can be elicitated in a human-machine setup by means of Wizard of Oz techniques, where an artificial agent is operated by a human operator without the subject knowing (McKeown et al., 2010), or in a human-human interactions by capturing TV recordings (Grimm et al., 2008).

Data processing methods for emotional speech are, of course, a hot research topic in these years. Both features and classification methods are being heavily tested in order to find a minimal set of features and a classification schema. Finding a general representation of emotional speech and a universally efficient classification method has been challenging until now. Recently, there has been a tendency in relying on automatic feature selection starting from very large acoustic feature sets (often more than 1000). This, in my opinion, is too much a *brute force* approach and is not as helpful in *understanding* the way emotions are conveyed as methods based on interpretable features set are. Other than acoustic features, classification a number of generic and specifically designed approaches have been proposed.

Support Vector Machines (SVMs) (Vapnik, 1982) and Regressors (SVRs) (Vapnik, 1995) are considered to be state-of-the-art among general purpose classifiers respectively for the discrete classification and continuous regression tasks. Linguistically motivated models coming from research on prosody have also inspired recent models like the hierarcical graphical model used in Fernandez and Picard (2011), which combines local information, based on syllables, and global information, based on global statistics over the acoustic properties of the utterance.

1.3.2 Emotional robotics

Emotions are a strong force influencing the way humans choose which actions to employ to correctly respond to arising situations. Responding correctly is not limited to survival related tasks but also to social conventions: acting emotionally, under the constraint of timing and modulation, significantly affects how people are perceived and how much easy is for them to acquire high social standings. At the same time, emotions continue to assolve their primordial goal of physically preparing the body to react to external stimuli. The physiological component of emotions often led to criticism about the attempt of introducing emotions into automated systems. Doubts arise both when it comes to the challenge of making computers and robots *feel* emotions and when it comes to introducing emotional communication. In (James, 1884, p. 190) it was stated that

> Without the bodily states following on the perception, the latter would be purely cognitive in form, pale, colourless, destitute of emotional warmth.

In the same paper, the author also argued (James, 1884, p. 193) that

> Emotion dissociated from all bodily feeling is inconceivable.

Given the specificity of the emotional experience from a physiological point of view for the (organic) human body, we should ask ourselves if it possible to find a correspondence in the (electro-mechanical) body of a robot. Should this not be possible, we would lose an

important component of human emotions seriously damaging the credibility of anything we would dare to call a synthetic emotion. Obviously, robots do not have to regulate, for example, blood pressure because they do not need muscles preparation to produce a particularly strong effort. Nevertheless, we can refer to the needs of a robotic body, different from from the ones organic bodies have, to look for an electronic counterpart of the physiological experience creating emotions.

To employ emotions in a robotic system, it is therefore necessary to take into account the two levels on which emotions operate: the cognitive level, which mainly involves decision making processes, and the physiological level, which is related to the way the robot can efficiently take advantage of its sensorial apparatus.

There have been multiple attempts to mimic the psychological effects of emotions in decision making processes because of the applications in software systems like virtual conversational agents (Becker et al., 2004; Bevacqua et al., 2010). In robotics, emotionally influenced planning of action sequences was presented in Kim and Kwon (2010); Gordon et al. (2010). These works concentrated on modeling the influence of the emotional experience by means of the appraisal theory summarized in Section 1.2.3.

In Kim and Kwon (2010), for example, the Kaist Motion Expressive Robot (KaMERo) was presented. Experiments with KaMERo were designed to assess how much easier is for human people to interact with a robot that expresses emotions in a multimodal way. KaMERo was programmed to play the Game of Twenty-Questions and it was equipped with touch sensors, face recognition and voice recognition capabilities to establish a deeply interactive experience with the user. KaMERo would have shown happy cartoonish faces while playing nice sound effects in relation to the answers given by the human player. The system was therefore composed of a logical module which tried to identify which question was the most likely to help the robot win the game and of an emotional module which influeced the robot's behavior with respect to the user. As it is shown in Figure 1.5, the planning module, implemented as a Partially Observable Markov Decision Process (POMDP) was coupled with a Deliberative Emotion Generation System (DEGS) to produce logically founded behaviors modulated by emotional reactions. Roseman's cognitive

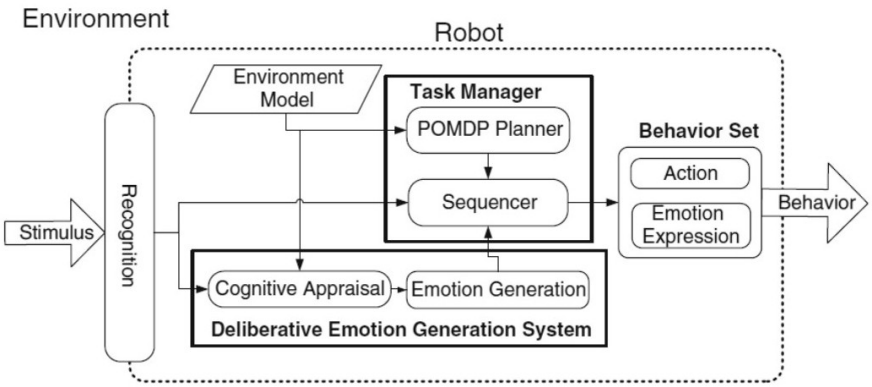


Figure 1.5: The cognitive appraisal architecture used in KaMERo

appraisal processes (see 1.4) were employed to generate KaMERo's emotional behavior. In the proposed architecture, a set of standalone algorithms was developed to produce a probabilistic interpretation of the result of each appraisal process.

Other than applications emotions can have in a cognitive architecture, the physiological interpretation of emotions has also inspired self-regulation applications in robotics that are motivated by neurobiology and related to attentional mechanisms. These can be designed both to direct processing power towards *arousing* stimuli or to introduce asynchronies in the periodic activation of behaviors in a robot in order to obtain an automatic adaptation in the emergent behavior without having an explicit action selection mechanism Burattini and Rossi (2010). This last work, in particular, shows that explicit cognitive processes are not necessarily needed to select actions on a low conscience level.

1.4 Neurobiology, psychology and robotics

In this work, I will concentrate on a linguistically motivated speech processing method for dimensional and continuous emotional tracking and on the definition of a robotics architecture to simulate a simple emotional intelligence without recurring to higher cognitive processes. In my opinion, higher functions should be included in an artificial intelligence only after exploiting as much as possibile on the lower levels of pure wired and instinctive

behavior. We should, therefore, check how much *intelligence* can be perceived in an artifact working on the basis on synthetic emotions before concentrating on the definition of more complex processes to integrate what will be missing. As it was pointed out in the very influential work by (Brooks, 1990, p. 13)

> "It is unfair to claim that an elephant has no intelligence worth studying just because it does not play chess."

In my view, disregarding lower levels of intelligence is not just unfair but totally wrong because of the principle of least effort, defined as follows (Zipf, 1949, p. 1)

> "In simple terms, the Principle of Least Effort means, for example, that a person in solving his immediate problems will view these against the background of his future problems, as estimated by himself. Moreover, he will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems. That in turn means that the person will strive to minimize the probable average rate of his work-expenditure (over time). And in so doing he will be minimizing his effort. Least effort, therefore, is a variant of least work."

This principle has been applied to a wide range of situations concerning not only human behavior, but animal behavior in general. For example, this tendency to economicity has been observed for reduction phenomena in speech minimizing articulatory effort (Millward, 1996), on information seeking in general (Poole, 1985) and in particular concerning the way people use websites (A. and A., 2002).

By considering the importance this *energy saving* principle has in nature, it appears to be logical to structure an approach to artificial intelligence design by first considering what it can be accomplished by the lower levels of intelligence, less powerful but also less demanding from the required effort point of view. In the framework described by Zipf (1949), if it is possible to spend less effort to solve an immediate problem by using emotional or purely wired systems, these *must* be used. In fact, preserving energy may

increase the probability of the system being able to solve more future problems. Higher level capabilities like symbolic analysis and long range planning, more powerful but more expensive in terms of energy and time spent, should be introduced as an integration to these basic processes rather than being considered the main resource the human brain relies onto. That is, in my view emotional processes are mainstream in brain activity while symbolic reasoning and problem solving capabilities represent auxiliary processes that are *ignited* by emotions. Of course, rules concerning how and when to invoke higher cognitive levels should also be researched.

In Chapter 4, I will follow the *distributed* view of emotions described in Ledoux (1998) both in a *vertical* and in a *horizontal* sense. Strictly following Ledoux's view of how perceptual systems contribute in a different and parallel way in defining the emotional experience, I define a *horizontal* dimension in which sensors and data processing modules contribute to the emotion definition. These contributions are then *vertically* organized by means of the dimensional model proposed by Fontaine et al. (2007) to abstract the emotional response on which reactions are based. The proposed architecture, while taking into account neurobiology on the horizontal dimension, considers emotions as a much more tangible and localized object by introducing a dedicated emotional model on the vertical dimension. This is because, while neurobiology suggests an efficient organization design for what it concerns perception and data flow evaluation in an emotional sense, the linguistic trick represented by the term *emotion* give an efficient and synthetic representation of the results of the emotional processing. By using the dimensional model, it is possible to avoid the limits of a discrete set of *emotional words* and obtain an efficient interface between perception and action. This interface is abstract enough to collect and generalize the results coming from a distributed emotional evaluation by decomposing the emotional effect into cross-culturally valid components. On the other hand, the interface is practical enough to let designers associate behaviors to the emotional states by exploiting the common way of referring to emotions as a whole. That is, the interface lets data processing modules contribute to the emotional state in a biologically motivated way and it abstracts the results into an unnamed *emotional state* representing a familiar basis on which designers can define

behaviors. In Figure 1.6, I summarize the contributions neurobiology and psychology give, in my view, to the definition of a synthetic emotion. While neurobiology inspires the organization of the contributions of the data coming from sensors that do not necessarily mimic human perception capabilities, psychological models organize the results of the separated emotional evaluation processes into a single object, called synthetic emotion. While this object does not exist in the human brain, as specified by Ledoux (1998), it provides a convenient basis for behavioral design as it describes a complex set of emotional responses in simple terms by summarizing multiple evaluation processes and by abstracting the outcome into intuitive continuous components.

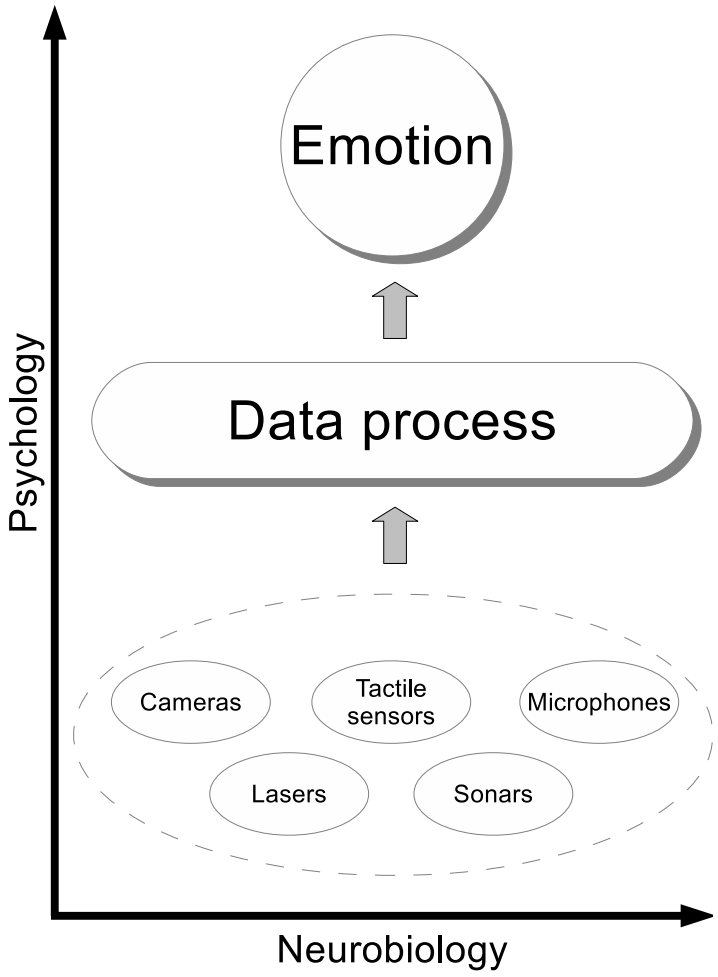


Figure 1.6: Neurobiology and psychology contribution to the definition of synthetic emotions in a robot.

Chapter 2

Speech processing with phonetic syllables

It is common to find, in the literature, approaches to speech processing for intonational features extraction that concentrate on global statistics of the pitch curve to obtain prosodic descriptors. While pitch is indeed the main correlate of intonation, the work of prosodists often consists in describing intonation phenomena in terms of the synchronization between pitch movements and segmental events like the occurrence of syllabic nuclei and boundaries. By considering global statistics, moreover, every part of the speech signal has the same importance of the others, contrasting with the literature concerning the perceptual phenomenon called prominence. In this Chapter, I will present a number of experiments on syllable-based speech processing concerning automatic syllabification, pitch stylization and prominence detection. The results presented in this Chapter will be the basis of features extraction technique presented in the next one.

2.1 Prosody

The term *prosody* comes from the greek $\pi\rho o\sigma\omega\delta\acute{\iota}\alpha$, which is composed by the two words $\pi\rho o\varsigma$ (near) and $\omega\delta\acute{\eta}$ (song). It refers to all the melodical, qualitative and rhythmical

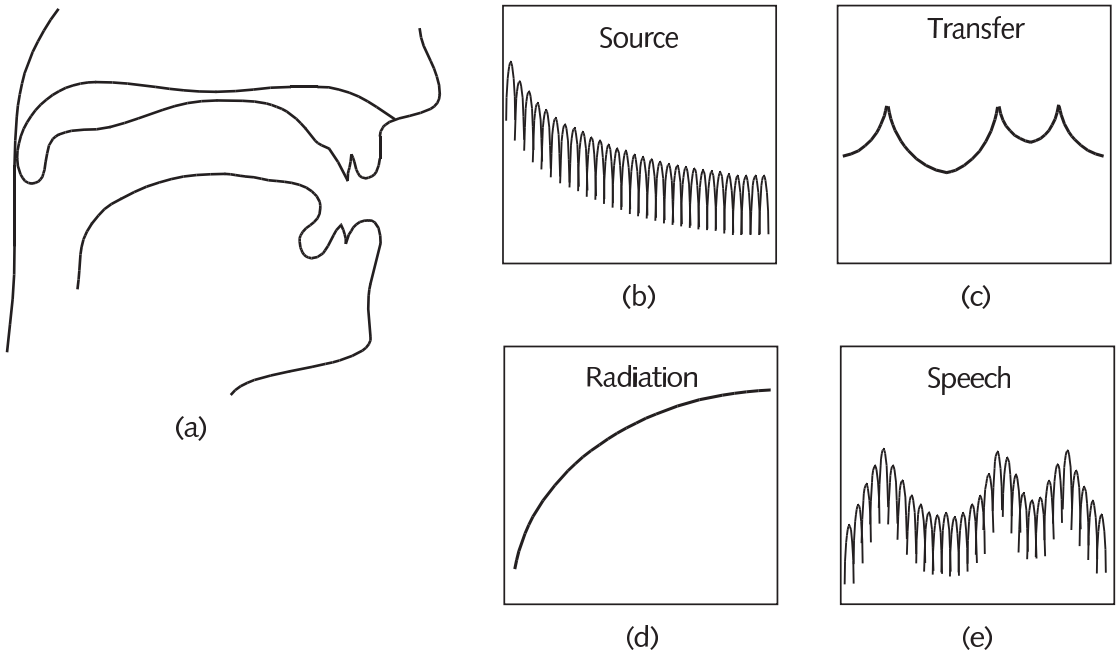


Figure 2.1: The source-filter model as described in Fant (1960). (a) represents a cross-section of the human vocal tract, (b) shows the spectrum produced by the vocal folds while vibrating, (c) shows the resonance spectrum of the vocal tract when it is configured to produce the /æ/ vowel, (d) shows the effect applied on the sound after radiating out of the vocal tract to result in the final spectrum (d)

components of speech that enrich the semantic content, often critically for what it concerns the correct transmission of a message.

First of all, it is important to consider the phonatory apparatus and how humans use it. These sounds, in general, are produced by letting the airflow coming from the lungs resonate in the vocal tract. The configuration of the vocal tract can be described in articulatory terms, by considering the position of articulators like tongue and lips, and in spectral terms, by considering the enhancing and dampening effects different part of the vocal tract applied on specific frequencies of a basic signal coming from the vocal folds. The spectral view of voice production is known as the source-filter model (Fant, 1960) and is summarized in Figure 2.1.

Depending on whether the vocal folds vibrate or are kept open while the airflow is passing, speech can be respectively voiced or unvoiced. While the vocal folds vibrate, the source signal is periodic and the frequency at which the vocal folds open and close consti-

tutes the fundamental frequency ($F0$) of a voiced sound. $F0$ is indeed the main correlate of intonation but, to study the role intonation has in communication, it is necessary to consider the integration phenomena the human ear and brain apply to the received signal before decoding it. For example, it is well known in the literature that the human ear is less capable of discerning high frequencies than low frequencies. The most widely accepted explanation, in psychoacoustics, about this is the *place theory* (Von Bekesy, 1960). This theory relates frequency perception to the area of the basilar membrane that resonates together with an submitted stimulus, activating the connected neural terminations (hair cells). The perceptual correlate of $F0$ is *pitch* and it is usually computed by means of a logarithmic transformation of the observed $F0$ value to mimic the lower discriminative capability of the human ear at higher frequencies. Of course, pitch varies as a function of time during speech production so pitch movements are a fundamental part of prosodic analysis.

Pitch and pitch movements alone, however, are not sufficient to describe intonational patterns. In the literature, these cues are always considered in terms of their relationship with the segmental level, on which elements like syllables, words and phrases are realized. Since intonational events are typically described as *anchored* to specific segmental events, the level on which they are realized is called supra-segmental. The autosegmental-metrical theory to intonational phonology (Pierrehumbert, 1980; Ladd, 1996) is the most important example of this close relation.

The synchronization between pitch movements and the occurrence of segmental material is very subtle and, even when small shifts are introduced, the human ear is able to detect them, possibly assigning a completely different meaning to the utterance. A very clear example of this is reported in D'Imperio and House (1997). In that work, the authors show that, by altering the alignment of a peak tone with the occurrence of a stressed vowel, neapolitan native speakers associate a declarative or interrogative function to the utterance depending on the position of the peak with respect to the syllabic nucleus. Experiments were performed both by altering an originally declarative production and an originally interrogative production. In order to analyze the correlation of the utterance's

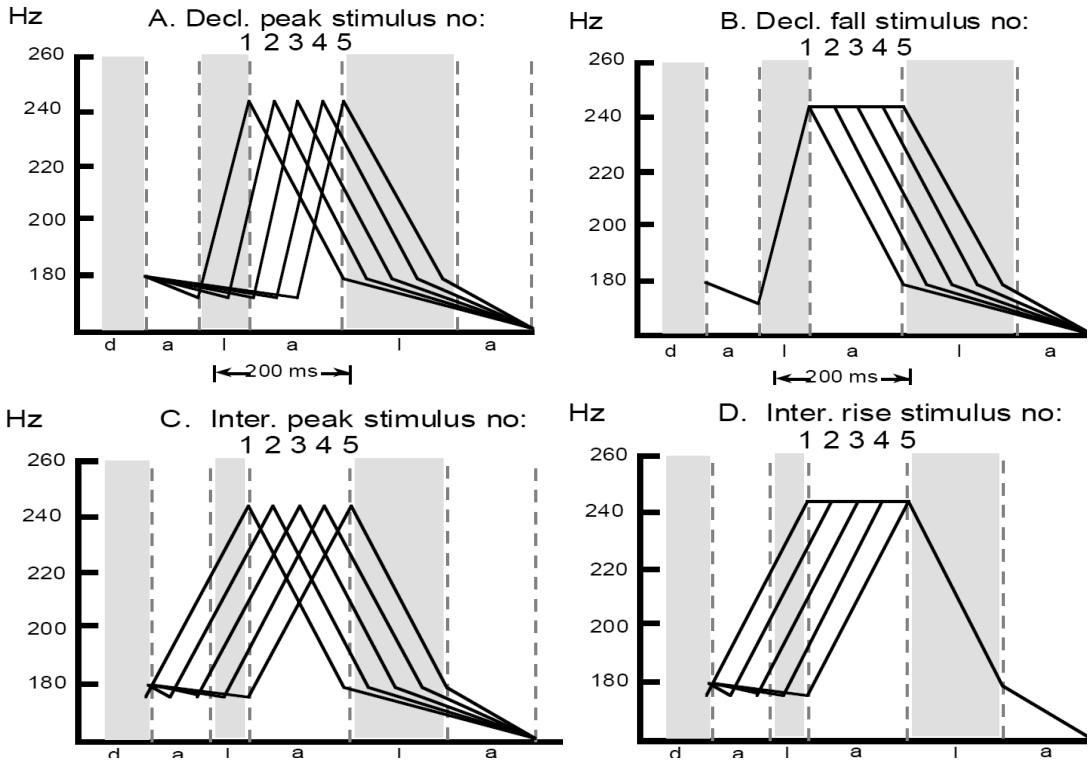


Figure 2.2: The stimuli used in D'Imperio and House (1997)

function both with the temporal alignment of the tonal peak and with the the presence of pitch movements inside the vowel, the artificial utterances (synthesized with the PSOLA algorithm) were produced by altering both the tonal peak alignment and the movement occurring inside the vowel. For each type of alterations, 4 shifting steps of equal duration were applied. Figure 2.2 shows a summary of the stimuli used in the experiment.

The number of *equal* responses obtained by human judges, summarized in Figure 2.3, shows that, as alterations get more significant, the category shift becomes more evident, especially for stimuli in which the tonal peak alignment was altered.

This study shows very clearly how the message conveyed by intonation can be radically changed as the synchronization between the segmental elements and the suprasegmental elements is altered. Moreover, it should be noted that stressed syllables usually constitute the area on which prosodists attention concentrates and that, as intonational strategies can be realized only in presence of voiced sounds, vowels have a particularly important role

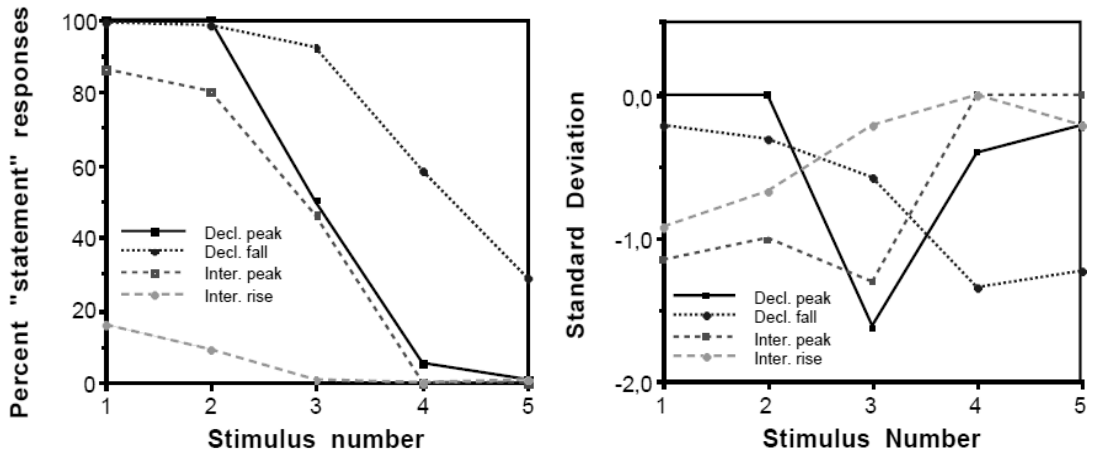


Figure 2.3: Graphical summary of the results presented in D'Imperio and House (1997)

in prosodic analysis. This indicates that, when analyzing a speech utterance, automatic systems should not consider all areas of the associated signal as equally important.

This example makes it clear that, to obtain a linguistically motivated representation of prosody to be used in a technological framework, it is necessary to introduce an analysis method that takes into account the same elements considered by linguists to build their theories and frameworks. Specifically for the the technological framework I am presenting here, I will take into account

- the basic segmental units used in linguistics to describe prosodic phenomena
- a perceptually consistent description of the pitch curve
- the relative weight each unit has with respect to its neighbouring units

2.2 Syllabification

Syllable segmentation is important in speech processing because it is connected with the main prosodic factors including rhythm and tempo and also because the opinion that syllables can be used as basic units in speech recognition has been investigated for a long time, see for example Wu et al. (1997); Jones et al. (1997). At the same time, the definition

of *syllable* is still controversial. It depends, among others, on the observed language, on the phonotactical rules involved in the morpho-phonological description adopted for that language and on some particular phonetic constraints. Moreover, as highlighted in (Sawusch, 2005, p. 7),

> "While descriptions of language and language processes use terms like word, phrase, syllable, intonation, and phoneme, it is important to remember that these are explanatory constructs and not observable events. The observable events are the movements of the articulators and the resulting sound. Consequently, understanding the nature of speech sounds is critical to understanding both the mental processes of production and perception."

In this work we are interested in which acoustic cues can be useful for an automatic syllabic segmentation. In the field of articulatory phonetics and phonology some authors link syllables with jaw movement (De Saussure, 1967), some others to chest burst (Stetson, 1951) or they consider syllables as the basic units of speech programming (Kozhevnikov and Chistovich, 1966). From the acoustic point of view, energy temporal patterns play a fundamental role: Jespersen (1920) was the first one to link syllabification with energy oscillation, observing that syllable nuclei are usually found in correspondence with energy maxima, while syllable boundaries correlate with energy minima. A first attempt to automatically segment a speech utterance into syllabic portions was presented in Mermelstein (1975). In this work a loudness function obtained by assigning a weight to each element within a set of spectral bands was used. An algorithm evaluating the shape of the loudness pattern (convex-hull) was then employed to find syllable boundaries.

In Pfitzinger et al. (1996), the speech signal was processed in three steps: first the authors used a bandpass filter, then they computed the energy pattern using a short term window and finally they low-pass filtered this energy function. The syllable nuclei were found by searching the local maxima of the energy contour. Another important result of Pfitzinger and colleagues was the comparison of the different manual syllabic segmentation that were done by several human labelers. They found an agreement of only 96% on

nuclei positions, making them assume this value as an upper bound for any automatic segmentation.

Another approach for speech syllabification was proposed by Jittiwarangkul et al. (1998). Their method was based on energy computation and successive smoothing. They tested various kinds of temporal energy functions for syllable boundary detection. The behavior of their algorithm depended on a number of empirically predefined thresholds.

In Reichl and Ruske (1993) one of the first attempts to use neural networks to segment speech into syllables was presented. The goal of this work was to find syllabic nuclei in German read sentences. The features extracted from the speech signal and given as input to the network consisted of Bark-scaled loudness spectra that were calculated every 10 ms. Two kinds of artificial neural networks were compared: a multilayer perceptron and a radial basis function neural network.

In Wu et al. (1997) the analysis method was based on smoothed speech spectra computed by two dimensional filtering techniques. This way the energy changes of the order of 150ms were enhanced while other techniques to emphasize the syllable onsets were used. The average energy over nine critical frequency bands every 10ms was also considered. The resulting vector was concatenated with log-RASTA features and was provided as input for a multilayer perceptron.

In Greenberg and Kingsbury (1997) the speech modulation spectrogram, a system for searching invariant features related to frequency portions of the speech spectrum, distributed across critical band-like channels, was introduced. According to Greenberg, invariants are mainly positioned in slowly varying dynamic features of the speech signal. The processing and recognition of speech features involves temporal constants that take two kinds of factors into account: speech rhythm parameters and the auditory temporal integration of the slowest spectral components.

Starting from the modulation spectrogram, a different kind of neural network was used in Shastri et al. (1999), specifically the temporal flow network that was previously introduced in Watrous (1993). With this tool the authors computed a function having local peaks at syllabic nuclei. The main differences between this net and the multilayer

perceptron is that the employed one allows recurrent links and time delay.

In Nagarajan et al. (2003) an automatic syllable segmenter using the minimum phase group delay function was developed. The author's approach is deterministic in the sense that they don't make use of stochastic evaluations about the signal. In their work they try to face the principal problem of the classic approaches to segmentation using the short term energy function, that is thresholding and energy fluctuations. If we consider the short term energy function as a magnitude spectrum, it can be demonstrated that it is associated to a minimum phase signal. The study of the negative derivative of the short term energy function (that is the "group delay function", if it was a magnitude spectrum) shows that it has peaks at syllable boundaries which are less sensible to energy fluctuations. This approach tries to find a more reliable reference to establish a decision threshold for syllable boundaries. An error rate of utmost 40ms for the 83% of the syllable segments suggests that this is one of the most powerful approaches found in literature. Continuation of this work was also presented in Prasad et al. (2004).

In Petrillo and Cutugno (2003) an algorithm employing energy analysis to set syllable boundaries corresponding to energy minima between two maxima was presented. Additional strategies to refine the initial result were employed to avoid segments containing fricative sounds only and to recompact long stressed vowels that were erroneously splitted. The values used for the set of parameters needed to perform automatic syllabification were obtained by using a number of function minimization techniques like genetic programming (Carnahan and Sinha, 2001) and simulated annealing (Kirpatrick et al., 1983).

In this Section I will describe a syllabification approach developed on the basis of the algorithm presented by Petrillo and Cutugno (2003).

2.2.1 Energy profile extraction

Since we are interested in detecting syllable nuclei, which are mainly vowel-like sounds, we filter the input signal to remove all the irrelevant spectral data like, for example, fricative noise. This filtering step makes energy peaks caused by the occurrence of a syllable nucleus

to stand out better than it would have been by looking at the energy profile extracted from the raw signal. For the presented approach, a band-pass filter with a lower cutoff frequency, set at 150 Hz, and an upper cutoff frequency set at 2800 Hz is used.

It is also necessary to smooth the energy profile in order to remove a high number of very weak energy peaks that would be taken into account as syllable nuclei candidates. Since we know that this kind of energy peak will never correspond to an actual syllable nucleus, it is necessary to avoid having the algorithm considering them as candidates by employing smoothing. This is done using the built-in PRAAT energy profile extraction procedure: the values in the sound object are first squared and then convolved with a Gaussian analysis window (Kaiser-20, sidelobes below -190 dB). Since the effective duration of the analysis window is computed as the ratio between 3.2 and the minimum periodicity frequency in the signal, which can be set by the user, we can obtain a smoother energy contour by lowering the minimum pitch parameter. In our experiments, we set the minimum pitch to 80 Hz.

In Figure 2.4 the energy profile of a raw speech signal along with its spectrum and its manual segmentation into syllable units is shown while in Figure 2.5 the smoothed energy profile and the filtered spectrum of the same speech signal is shown. In the example it is possible to see how this preprocessing step removes many energy peaks caused by phenomena other than syllable nuclei occurrence.

2.2.2 Syllable nuclei candidates detection

After filtering the speech signal in the frequency domain and obtaining a clean energy profile, the algorithm applies filtering in the time domain to remove artifact peaks that are not to be considered syllable nuclei candidates. It builds a syllable nuclei candidate list that, in the beginning, contains all the energy peaks that were not removed during the previous step. Then, it removes from the list the energy peaks that do not appear to be eligible as syllable nuclei candidates according to a second analysis step.

First of all, silent areas in the speech signal are computed using a silence threshold

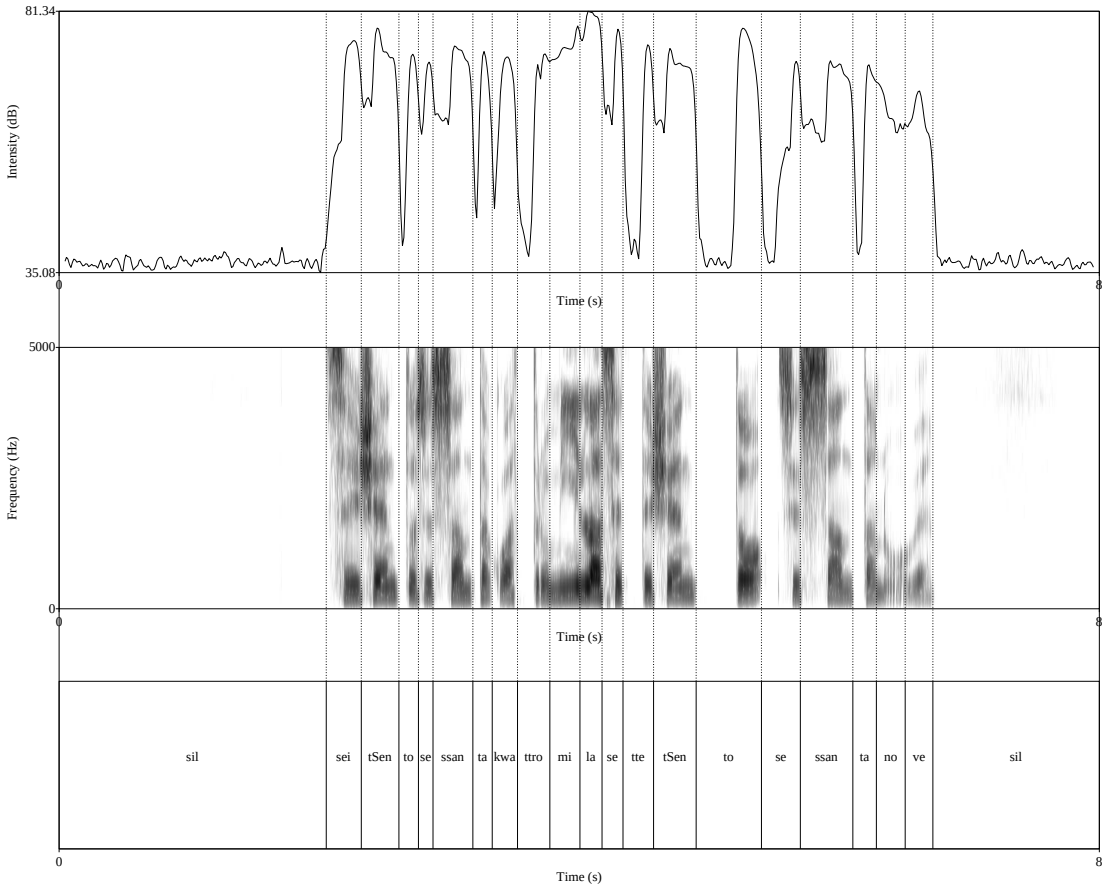


Figure 2.4: The original signal along with its manual syllabic segmentation. The first frame shows the energy profile while the second one shows the unfiltered spectrum.

defined according to the PRAAT standard. Areas where the energy value is less than the maximum energy value minus a user defined parameter (which we set to 30 db) are considered silent. Peaks falling in silent areas of the speech signal are removed from the candidate list.

Filtering in the time domain is especially needed to remove a particular kind of artifact energy peak that is resistant to frequency domain filtering. The artifact we want to remove during this step shows up as a weak energy peak on a steep rise of the intensity profile. This is mainly caused by the occurrence of alveolar trills, which are very frequent in Italian, and is particulary hard to avoid during syllable nuclei candidates evaluation since its phonation type is voiced. In Figure 2.6 an example of an artifact peak caused by the alveolar trill

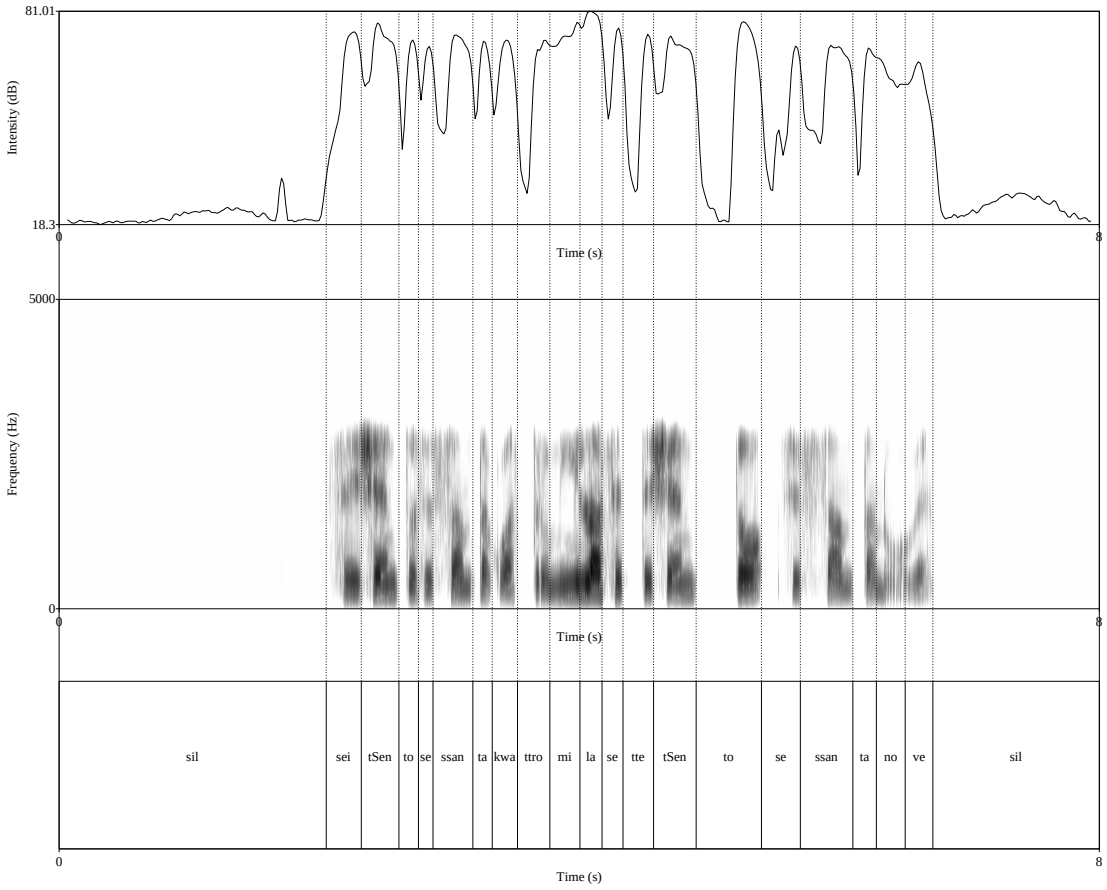


Figure 2.5: The filtered signal along with its manual syllabic segmentation. The first frame shows the smoothed energy profile while the second one shows the filtered spectrum.

occurring inside the [tre] syllable is shown.

In order to avoid the systematic insertion errors caused by this artifact, the algorithm uses a template to detect this specific situation. This procedure scans the syllable nuclei candidate list searching for peaks showing up on an energy rising that reaches its top in, at most, 100 ms. If the difference between the value of the peak and the value of its following energy minimum is less than 15% of the total rise, the peak is recognized as artifact and removed from the candidate list. When an artifact is removed, its immediately following energy dip is removed from the list of energy minima too. This is because by removing an energy peak the subsequent valley becomes inconsistent and must not be taken into account anymore when evaluating the next energy peak.

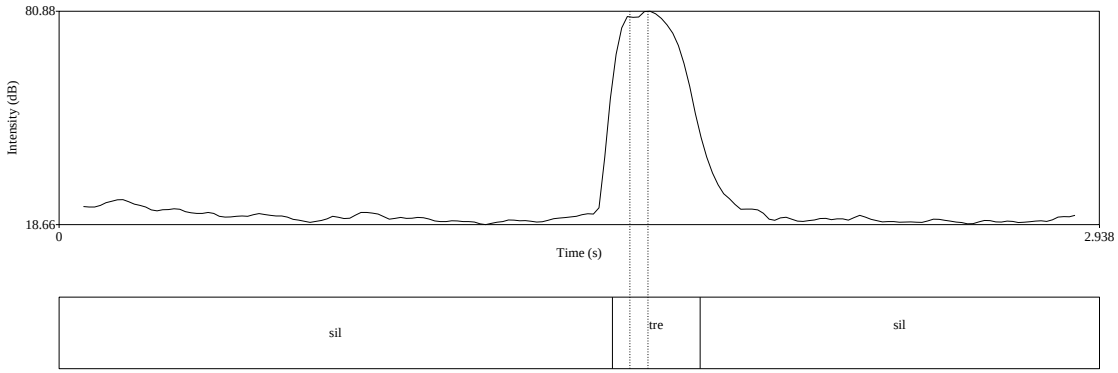


Figure 2.6: An artifact peak caused by an alveolar trill. This artifact, if not detected, causes an insertion error because of a false positive occurrence during the syllable nuclei detection step.

2.2.3 Syllable boundary markers positioning

Having found syllable nuclei, the algorithm needs to set syllabic boundaries. A specialized strategy is employed for the first and the last syllable marker because their position is determined by the silence threshold being crossed. In particular, we found that the position of the first marker must be adjusted when the sentence starts with a fricative consonant. To correctly position the first marker, we employed the same approach used in Petrillo and Cutugno (2003) to set the generic syllable marker. First, we position the marker where the silence threshold is exceeded and then we compute the *residual energy* by low pass filtering the signal using a cutoff value of 1100Hz. We left-shift the marker until the difference between the original signal and the filtered one is inferior to 1db. This strategy leaves the marker near to the point in which the silence threshold is exceeded when there is not a fricative consonant while recovering it when it occurs at the beginning of the sentence.

Regarding the other markers, for each found syllable nucleus, the preceding energy dip is considered as the initial position of the related marker. The marker is then moved by considering the following

$$t_b = argmin_{t \in [t_p; t_d]} E'(t) \tag{2.1}$$

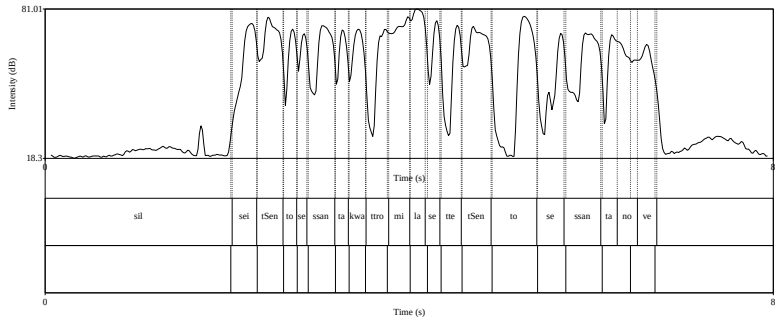


Figure 2.7: The energy profile of a speech signal along with its manual and automatic syllabification

where t_b is the time at which the syllable boundary is finally positioned, t_p is the time at which the preceding nucleus is found, t_d is the time of the original position of the boundary and E is the energy profile of the signal. Through Equation (2.1), we position the syllable marker where the slope of the energy profile between the preceding nucleus and the original position of the marker itself reaches its maximum. The last step the algorithm performs is to merge syllables that are shorter than a user defined threshold (we used 70 ms) with the syllable on their right side. The last syllable, if too short, is merged with the one at its left side.

In Figure 2.7 we show the final result of the syllabification procedure compared to the manual one.

2.2.4 Evaluation

In order to evaluate the automatic segmentation procedure, it is necessary to compare the output of the system with manually annotated boundaries. The evaluation of automatic analysis is expressed in terms of the number of differences between automatic and human analysis. These differences, here called errors, can be of three types:

- Deletions
- Insertions
- Substitutions

Figure 2.8: A very distant marker being considered a substitution because of the search region being too large

A deletion error occurs when a syllable is not recognized at all, i.e. when two syllables are expected and only one segment is found. Insertion errors, on the opposite, occur when an expected syllable is split into two segments. The last kind of error produces a difference between the temporal positions of the separation marker from the two types of annotation that results greater than a fixed threshold, without altering the number of recognized segments.

The testing protocol is a modified version of the system presented by Petek et al. (1996) and is documented in Ludusan (2010). Consider $R = (r_1, ..., r_n)$ and $S = (s_1, ..., s_m)$ the string of the manual segmentation and the string of the automatic segmentation. A search region for each element r_i (with $i = 1...n$) is defined and all the elements s_j are assigned to their corresponding region. If no marker s_j can be found in a certain search region, the marker corresponding to r_i is considered a deletion. If there is just one marker s_j in a search region, it is a substitution of r_i. If there are two or more markers in a search region, the nearest is considered a substitution while all the others are considered insertions.

As suggested by Petek et al. (1996), a search region of the marker r_i starts at half of the interval $[r_{i-1}, r_i]$ and it ends at half of the interval $[r_i, r_{i+1}]$, thus excluding the possibility of overlapping regions.

This way, however, search regions may become too large or too small. In case of large regions it may be possible to consider as a substitution a marker quite distant from the reference marker. This case is illustrated in Figure 2.8.

In this example, in the search region of the marker r_{i+1} there is only s_{j+1}, so the algorithm will consider s_{j+1} as if it was a substitution of r_{i+1}. Since the two markers

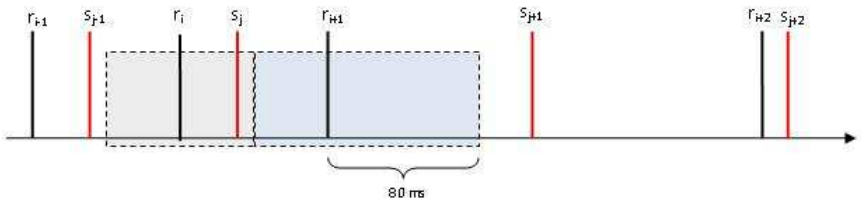


Figure 2.9: Thresholding to limit the size of each semi-regions solves the problem of distant markers being considered as substitutions.

Figure 2.10: An example of an insertion-deletion couple being inserted in place of a substitution because of the search region being too small.

are very distant from each other, however, it is better to consider them as an insertion and a deletion. To solve this problem a threshold to limit the size of each semi-region is introduced (see Figure 2.9).

On the contrary, if the search region is too small, an insertion and a deletion are introduced instead of a substitution. In the example shown in Figure 2.10, s_j is considered an insertion and r_{i+1} a deletion but they are close enough to consider s_j a substitution of r_{i+1}. To solve this, all the consecutive pairs (insertion-deletion), or viceversa, are processed and they are converted into substitutions, depending on the distance between the two markers.

As baseline for this test, the algorithm presented in Petrillo and Cutugno (2003) is considered. In Table 2.1 results obtained by the two approaches with the employed evaluation system are summarized.

	Substitutions	Insertions	Deletions	Correctness	Accuracy
Proposed algorithm	2.03	6.19	5.72	91.74	85.14
Petrillo-Cutugno	4.13	5.74	5.61	89.67	83.58

Table 2.1: Results obtained on the SPEECON corpus (in %) by the new algorithm and by the baseline approach for Italian.

2.3 Pitch stylization: preliminar analysis

The modulation of pitch plays a prominent role in everyday communication fulfilling very different functions, like contributing to the segmentation of speech into syntactic and informational units, specifying the modality of the sentence, regulating the speaker-listener interaction, expressing the attitudinal and emotional state of the speaker, and many others (see Vassière (2005) for a complete overview). It is therefore not surprising that research on pitch has received great attention in recent years. However, both the phonetic description of pitch movements and their communicative interpretation still present several methodological and theoretical challenges. I will concentrate only on the task of modeling the pitch curve on a psycho-acoustical basis, thus adopting the principle that when analyzing communicative intents attention should be focused on what is heard rather than what is spoken. For what it specifically concerns pitch stylization, as it was pointed out in (t'Hart et al., 1990, p. 25)

> "[...] No matter how systematically a phenomenon may be found to occur through a visual inspection of F_0 curves, if it cannot be heard, it cannot play a part in communication."

Among the first attempts to follow this principle, it is important to highlight the work presented in (Hirst and Espesser, 1993) where the pitch curve was considered as the result of the composition of a micro-prosodic component, intended as perturbations caused by mere articulation, and of a macro-prosodic one. The MOMEL algorithm (Hirst and Espesser, 1993; Hirst et al., 2000) was designed to separate the micro-prosodic component from

the macro-prosodic one, which was represented by means of a quadratic spline function. The MOMEL algorithm is often used as a term of comparison for stylization algorithms although it was designed to produce a *model* of macro-melody rather than a stylization. The difference lies in the fact that the output of the MOMEL algorithm does not discard the micro-melodic component, thus allowing to recover the original pitch profile. Being widely used in prosodic research, however, the macro-melodic component of the MOMEL output constitutes a reference for all representations of prosody aiming at describing general intonation profiles.

Research on pitch stylization algorithms is highly attractive both for speech technologists and prosodists. An economic and perceptually reliable stylization is a fundamental component for many technologies of voice, like speech recognition, speech synthesis, language identification, and speaker recognition. On the other hand, for prosodists, a perceptually based stylization of pitch constitutes a solid ground to define a set of descriptive units of intonation. For example, this kind of set could be used to develop automatic or semi-automatic systems of prosodic annotation (Campione et al., 2000; Mertens, 2004). Lastly, integrating a tonal perception model in the stylization algorithm has important implications for basic research on pitch perception: by submitting re-synthesized stimuli to the human ear in perceptual tests, it is possible to evaluate different models of tonal perception and to provide important clues for developing more reliable models.

In this Section, I will first show the results of a preliminar study to verify the impact syllabic prominence can have on the task of pitch stylization and I will also show that statistical closeness measures are not a good estimator for what it concerns the quality of the stylized curve. The algorithm used in this preliminar study uses a set of empirically determined parameters that are removed in the following Sections, where I will report on the development of an original approach based on a tonal perception model that will be used for emotional features extraction in the next Chapter.

2.3.1 An adaptive strategy for pitch stylization

Although pitch is the perceptual correlate of F_0 variations, the relationship between the perceived pitch and the F_0 curve is not trivial. First, many micro changes in the F_0 curve do not have a perceptual counterpart; these variations are therefore irrelevant for communication and must be filtered out from the actually perceivable events. For this reason, the need of a stylization that maintains only the perceptually relevant aspects of the F_0 curve arises. Following the definition of stylization given by (t'Hart et al., 1990, p. 42),the synthetic curve

> "[...] should eventually be auditorily indistinguishable from the resynthesized original"

and, moreover, it

> "[...] should meet the additional requirement that it must contain the smallest possible number of straight-line segments with which the desired perceptual equality can be achieved."

This definition was interpreted as an optimization process for the first time in Ghosh and Narayanan (2009). In that work a Dynamic Programming (DP) algorithm was designed to find the optimal balance between an empirically determined number of segments, based on the findings of Wang and Narayanan (2005), and the Mean Square Error of the stylized curve with respect to the original one. In this section, this same interpretation is followed but a *divide et impera* approach is used instead of dynamic programming and the stylized curve is constrained to contain the smallest number of control points rather than segments. From hereon, I will refer to this approach the Optimal Stylization (OpS) algorithm. The cost of the stylized curve is also explicitly taken into account during computation and for evaluation purposes by measuring the number of points used by the different algorithms, which is an aspect that has often been overlooked in previous works.

First of all, it is necessary to show that, for the pitch stylization problem, the optimal substructure property holds as this is a necessary condition for *divide et impera* and dynamic programming approaches to be applicable. We therefore have to prove the following

Theorem 1. *Given a pitch curve P and a stylization S of P, if S is optimal, its subcurves must be optimal too with respect to the pitch curve section they stylize.*

Suppose we have a generic function $\boldsymbol{F}$ that evaluates, with respect to the definition, how good a stylized curve S is, given the original pitch curve P. To indicate a point on a sampled pitch curve the following notation will be used:

$$p_x = (v_{s_x}, t_{s_x}) \tag{2.2}$$

where v_{p_x} is the value in semitones of the point and t_{s_x} is the time instant of the point in ms. Both P and S can therefore be described as sequences of points so that $P = [\mathbf{p}_1, ..., \mathbf{p}_n]$ and $S = [\mathbf{s}_1, ..., \mathbf{s}_m]$ with $m \leq n$. If S is an optimal stylization for P, then $\boldsymbol{F}(S) \geq \boldsymbol{F}(\overline{S})$ holds for every possible stylization $\overline{S}$ of P.

Given an index i such that $1 < i < m$, we consider the two subcurves $S' = [\mathbf{s}_1, ..., \mathbf{s}_i]$ and $S'' = [\mathbf{s}_i, ..., \mathbf{s}_m]$. We need to prove that, for every stylization $\overline{S'}$ and for every stylization $\overline{S''}$, both $\boldsymbol{F}(S') \geq \boldsymbol{F}(\overline{S'})$ and $\boldsymbol{F}(S'') \geq \boldsymbol{F}(\overline{S''})$ hold. Let us concentrate on S': if S' would not be optimal, some other $\overline{S'}$ should exist such that $\boldsymbol{F}(\overline{S'}) > \boldsymbol{F}(S')$. Since the $\boldsymbol{F}$ function evaluates the quality of the stylization on the basis of what it was stated by t'Hart et al. (1990), S' is either perceptually better and/or less expensive than S'. If there was such a curve, then it would be possible to replace S' with $\overline{S'}$ inside S, thus obtaining a new $\overline{S} \neq S$ such that $\boldsymbol{F}(\overline{S}) > \boldsymbol{F}(S)$, contradicting the initial hypothesis. Then, if S is optimal, its subcurves must be optimal too with respect to the pitch curve section they stylize (q.e.d.).

The DP algorithm presented by Ghosh and Narayanan (2009) was designed to optimize, in $O(KN^2)$, the Mean Square Error (MSE) of a pitch stylization by using a predetermined number K of segments for a curve with N segments. To automatically establish how many segments should be used to stylize the original pitch curve, the authors adopted an approach based on the results presented by Wang and Narayanan (2005). They performed

a multilevel decomposition of the pitch contour using a Daubechies wavelet (Db10) and considered the number of *extrema* in the third level of the decomposition to be equal to $K - 1$. The choice for the third level was motivated by the results of a perceptual test in which subjects indicated that the third decomposition level contained the optimal number of extrema in 60% of the cases (Wang and Narayanan, 2005). However, the same data also showed that in 27% of the cases the optimal decomposition was found in the fourth level, in 2% of the cases it was found in the fifth level and in 11% of the cases it was found in the first level. Since the higher the chosen level, the less the number of extrema, it can be concluded that, by systematically choosing the third level to estimate K, the number of segments in Ghosh and Narayanan (2009) was probably optimal in 60% of the cases, with 29% of the times being better to use fewer segments and 11% of the cases being better to use more segments.

The mathematical formulation of the pitch stylization problem presented here is more adherent to the definition given in t'Hart et al. (1990) than the one used by Ghosh and Narayanan (2009). Specifically, the goal of the optimization process should be to obtain the best balance between curve quality and cost. That is, the two constraints provided by the definition of stylization should be taken into account *at the same time* during computation. In the two-step process employed in Ghosh and Narayanan (2009), this was not the case and the DP algorithm was not able to correct a non-optimal choice of the K parameter because it was designed to find the minimum MSE stylization for a pre-determined number of segments. Since we have seen that the first step finds the optimal number of segments in most cases but not always, the second constraint is not guaranteed to be satisfied. In the formulation used here, $\boldsymbol{F}$ represents an evaluation of the balance between quality and cost of the stylized curve so that, given a quality function $q(S)$ and a cost function $c(S)$, we can describe $\boldsymbol{F}$ as a generic composition of the two measures. We therefore need to define the $q(S)$ and $c(S)$ functions.

In Ghosh and Narayanan (2009) it is assumed that MSE is the best measure to evaluate how similar the stylized curve will be *perceived* to be with respect to the original one. While assuming the same for now, one of the goals of this preliminar investigation is to check

if this is actually the best choice, given that perceptual phenomena are important in this task. Should this not be the case, it would be easy to substitute the $q(S)$ function with a more appropriate one while retaining the same framework. Considering Normalized Root Mean Square Error (NRMSE) as a quality measure we obtain

$$q(S) = 1 - \sqrt{\frac{\sum_{i=1}^{n} \left(\frac{\overline{p}_i - p_i}{\overline{p}_i} \right)^2}{n-2}} \tag{2.3}$$

where n is the number of points of the original curve, p_i is the i-th point of the original curve and $\overline{p}_i$ is the corresponding point on the stylized curve estimated after linear interpolation of the S curve control points. If we sample the original curve at a sufficiently small time interval (10ms in the presented tests) and we linearly interpolate it, we obtain a very good quality curve that will also be very expensive in terms of the number of points used. If we take this curve as reference, we can evaluate the cost of a stylization as the ratio between the number of points used in the proposed curve $|\overline{p}|$ and the number of points $|p|$ used in the reference one.

This ratio is weighted with a sigmoid function to evaluate the final cost parameter so that values of the cost measure at one end of the scale will not be very different. The value of the function $c(S)$ is therefore

$$c(S) = 1 - \left(\frac{1}{1 + e^{\frac{-(x-0.5)}{0.13})}} \right) \tag{2.4}$$

where $x = |\overline{p}|/|p|$. As $q(S)$ and $c(S)$ are two different performance measures of the same object, we can evaluate the balance between the two, which is $\boldsymbol{F}(S)$, by using the harmonic mean.

$$\boldsymbol{F}(S) = \frac{(1+\beta^2) q(S) c(S)}{\beta^2 q(S) + c(S)} \tag{2.5}$$

However, we would like to dynamically adjust how the two parameters are weighted by taking into account how complex the original curve seems to be. Specifically, we want to favor cost over quality in areas that can be approximated linearly while favoring quality

when important changes appear. In this experiment, this is accomplished by adjusting the β parameter on the basis of the standard deviation of the differences between consecutive points in the original curve and using it as a penalization factor to an initial value of 2, which means that cost is weighted twice quality. Therefore we have

$$\beta = 2 - stdev(\Delta(p_i, p_{i-1})) \tag{2.6}$$

We set the inferior limit of β to 0.5 for symmetry with respect to the effect of $\beta = 2$. This is an empyrical way of solving the problem that will be better addressed in Section 2.4.

Having set the $q(S)$ and $c(S)$ functions, we can design a *divide et impera* algorithm in which the original problem is iteratively split into two subproblems of approximately equal size until a subproblem with a trivial optimal solution is found. By backtracking and combining the solutions of each couple of subproblems an optimal solution for the original problem can be found. First, we define a function to obtain the possible merging of two curves sharing an endpoint:

$$g([\mathbf{s}_{a_1}, ..., \mathbf{s}_{a_k}], [\mathbf{s}_{b_1}, ..., \mathbf{s}_{b_z}]) = \left\{ \begin{array}{c} [\mathbf{s}_{a_1}, ..., \mathbf{s}_{a_{k-1}}, \mathbf{s}_{b_2}, ...\mathbf{s}_{b_z}], \\ [\mathbf{s}_{a_1}, ..., \mathbf{s}_{a_k}, \mathbf{s}_{b_2}, ..., \mathbf{s}_{b_z}] \end{array} \right\} \tag{2.7}$$

Then, we can define a recurrence relation as follows:

$$\left\{ \begin{array}{l} Opt([\mathbf{p}_1, \mathbf{p}_2]) = [\mathbf{p}_1, \mathbf{p}_2] \\ \\ Opt([\mathbf{p}_1, ..., \mathbf{p}_n]) = \underset{S}{\operatorname{argmax}} F(S) := \\ \left\{ \begin{array}{l} S \in g(Opt([\mathbf{p}_1, ..., \mathbf{p}_{mid}]), \\ Opt([\mathbf{p}_{mid}, ..., \mathbf{p}_n])) \end{array} \right\} \end{array} \right. \tag{2.8}$$

The procedure to solve the problem by means of this recurrence relation is shown by Algorithm 1. Computational complexity is $O(NlogN)$.

Algorithm 1 *The OpS algorithm*

procedure $merge(P, S_a, S_b)$

$k = length(S_a)$

$z = length(S_b)$

$f_{orig} = F(P, [s_{a_1}, ..., s_{a_k}, s_{b_2}, ..., s_{b_z}])$

$f_{mod} = F(P, [s_{a_1}, ..., s_{a_{k-1}}, s_{b_2}, ...s_{b_z}])$

if $f_{orig} <= f_{mod}$ **then**

return $[s_{a_1}, ..., s_{a_{k-1}}, s_{b_2}, ...s_{b_z}]$

return $[s_{a_1}, ..., s_{a_k}, s_{b_2}, ..., s_{b_z}]$

end procedure

procedure $optimize(P, p_{start}, p_{end})$

if $p_{end} - p_{start} < 2$ **then**

return $P[p_{start} : p_{end}]$

$p_{mid} = length(P[p_{start} : p_{end}])/2$

$S_a = optimize(P, p_{start}, p_{mid})$

$S_b = optimize(P, p_{mid}, p_{end})$

$S = merge(P[p_{start} : p_{end}], S_a, S_b)$

return S

end procedure

procedure $main(P)$

for all voiced segments P_v **do**

p_{start} = Get the P_v start point index

p_{end} = Get the P_v end point index

$S_v = optimize(P, p_{start}, p_{end})$

end procedure

One of the goals of this preliminar test is to evaluate the impact of syllabic prominence in the task of pitch stylization. I introduce here a variant using manual syllable level segmentation and prominent syllables annotation to obtain less expensive curves. The objective is to demonstrate the following

Hypothesis 1. *When stylizing a pitch contour, it is possible to use fewer points in pitch curve sections falling inside non-prominent syllables without damaging perceptual equality.*

In this variant of the OpS algorithm using manual annotations, which I will call OpSProm as opposed to OpSNoProm, we slightly modify the steepness of the sigmoid function employed by the cost function $c(S)$ to favor the removal of points in non-prominent syllables while we use the same cost function presented before to stylize the pitch curve inside prominent syllables. The reader is referred to Section 2.6 for more information on the automatic prominence detection problem and for the results of specific experiments performed on this task.

2.3.2 Testing methods

In order to take into account perceptual significance, the OpS algorithm was evaluated using both objective measures and a subjective listening test. The quality evaluation of the stylized curves proposed by OpS and by its variant was performed by comparing them with the ones proposed by the MOMEL algorithm and by the DP algorithm. The cost evaluation was performed on a larger corpus as well as the investigation of the effectiveness of statistical closeness as quality measure. This is because while the need to collect human subjective evaluations poses a limit on the number of samples to employ, the differences in the objective measures we wanted to observe were not definite enough to be captured by statistical tests on a limited number of samples.

2.3.3 Test material

For the objective evaluations the 382 files of the prominence annotated TIMIT subset used in Tamburini and Wagner (2007) were employed to test automatic methods for prominence detection. For the listening test, 20 sentences of duration varying between 2 and 3 seconds were selected from the CLIPS corpus of Italian semi-spontaneous speech (Savy and Cutugno, 2009). The chosen sentences contained a single tonal unit in order to obtain coherent intonational profiles for the listening test. The CLIPS subset was annotated by an expert linguist following the same method used for the TIMIT subset.

2.3.4 Listening test setup

For the subjective, qualitative, evaluation of the stylizations, a *humming track* of each original pitch curve and of each stylization was generated using the dedicated routine implemented in PRAAT. This was in order to allow the subjects to concentrate on the pitch curve and to be uninfluenced by semantic and pragmatic information. The listening test was designed following the approach presented in t'Hart et al. (1990) to evaluate the performance of each competing algorithm. Human listeners were presented with a series of stimuli pairs and were asked to judge if the stimuli in each pair were actually the same stimulus played two times or if they could detect any difference between the two. In each run of this test, 10 stimuli were chosen to be paired with those obtained by the used algorithm (group A), 5 further stimuli were paired with themselves (group B) and 5 further ones were paired with an intentionally altered version (group C). In these altered profiles, the pitch curve was shifted 10% up to introduce subtle but audible modifications. Mean pitch in the test files was 145Hz. Shifting the pitch curve 10% up means adding 14 Hz, thus obtaining new curves with 159Hz mean value. A relative difference of approximately 1.6 semitones between the original and the artificially altered curves was therefore introduced, falling in the region indicated by t'Hart (1981) (t'Hart et al., 1990, p.29) to indicate the just noticeable difference threshold in pitch, which was comprised between 1.5 and 2 semitones. This is different from what it was suggested by t'Hart et al. (1990): indications were of

shifting forward the pitch curve of about 50ms but it was not possible to do this in this particular experiment because, by using the humming track, this modification would have not introduced any difference detectable by the human ear: that is, the only difference would have been an initial silence 50ms longer as there would not have been any segmental content to contrast with. The stimuli assigned to each group were different among the four test runs to avoid the subjects to become acquainted with the stimuli of group C (the most easily recognizable ones). To limit the effect of tiredness and overtraining, each subject was presented the four runs in different order. Randomization for the presentation of the stimuli was also employed. 14 subjects (7 males and 7 females) were asked to evaluate if the paired humming tracks were equal or not. The (B+C) control group was used to evaluate the capability of the listeners to correctly discriminate equal stimuli (group B) from the ones in which clear differences were artificially introduced (group C).

The expected result of this test is that the proposed preliminar approaches are at least equivalent, from a qualitative point of view, to other approaches presented in the literature, thus validating Hypothesis 1.

2.3.5 Results

The OpS algorithm and its prominence based variant are compared here with the MOMEL algorithm and with the DP approach presented in Ghosh and Narayanan (2009). We evaluate the four approaches in terms of percentage of times that the proposed curves were judged to be equal by the listeners, in terms of points per second (Pps) used and in terms of NRMSE. In Table 2.2 the summary of the data is shown.

Table 2.2: Subjective (CLIPS) and objective (TIMIT) statistical comparison among the algorithms

	OpSNoProm	OpSProm	MOMEL	DP
Equality	72.86%	69.29%	76.43%	70.71%
Pps	3.78	3.47	3.75	5.72
NRMSE	0.1685	0.1975	0.1789	0.0663

A chi-squared test was used to compare the number of times the stylized curves were considered to be equal to the original one. The test found that the differences shown in Table 2.2 are never statistically significant ($p > 0.3$). The number of points used is therefore the key factor to decide how good the proposals are. An ANOVA test was used to evaluate the significance of the differences in terms of Pps. The test found that the OpSNoProm algorithm shows similar performance with respect to the MOMEL algorithm ($p > 0.5$). The OpSProm variant uses significantly fewer points with respect to the MOMEL algorithm ($p < 0.01$) while the DP approach clearly uses many more points than the others. Differences in terms of NRMSE were also evaluated with ANOVA and were always significant ($p < 0.01$). This means that although the DP approach, as expected, obtains the best NRMSE, it does not introduce an improvement in subjective evaluation, as shown by the chi-squared test.

Data presented in Table 2.2 show that no difference in perceptual equivalence is caused by the pps redction caused by the use of manual annotation of prominent syllables, validating Hypothesis 1. These results are presented in Origlia et al. (2011).

2.4 A tonal perception model for optimal pitch stylization

In this Section I will analyze strengths and weaknesses of the basic OpS approach in the light of the results summarized in the preceding section. The main updates to the original algorithm presented in this work are based on the observations reported here.

2.4.1 Observations

Observation 1. *The basic OpS algorithm equals the performance of the MOMEL algorithm with the additional advantage of being parameter independent.*

This observation comes from the fact that there was no statistically significant difference between the basic OpS algorithm and the reference MOMEL algorithm both on

the subjective test and on the number of points used. Observation 1 implicates that, for the same pitch curve, only one stylization can be found by the OpS algorithm while the MOMEL approach proposes a different solution each time it is provided with different parameters. Given the result of the ANOVA test on the NRMSE measures we can assume that the curves proposed by the two OpS variants and MOMEL can differ in shape up to a statistically detectable degree. However, the listening test, which is the ultimate way of verifying the quality of the curves in the pitch stylization task, highlights that these differences do not introduce appreciable differences to the human ear. This is in line with what was noted in (t'Hart et al., 1990, p. 42), where the term *close copy* is used as synonym of *stylization*

> "[...] there is not just one close copy for a given F_0 curve. The limits of dynamic pitch perception, together with restrictions of human memory capacity, make it possible that a second close copy would show small deviations if visually compared to the first one. Due to perceptual tolerances, however, they sound equal to each other, and to the resynthesized original."

Given the motivations stated in (Ghosh and Narayanan, 2009, p. 810) that

> "[...] optimality in terms of some objective function is necessary to understand the effect of parameterization of the pitch contour in a systematic way."

between two stylization approaches leading to equal perceptual performance, the one that yields a univocal solution is to be preferred at least for the standardization possibilities it offers to researchers working on prosody.

Observation 2. *Statistics about the S curve closeness to the P curve are not a good estimator of a stylized pitch curve quality.*

The chi-squared test performed on the subjective data shown in Table 2.2 and the ANOVA results on NRMSE statistics highlight that although the DP approach performed significantly better than the other approaches in terms of NRMSE, it did not introduce an

improvement in subjective evaluation. This is in line with the observation reported by the authors that (Ghosh and Narayanan, 2009, p. 813)

> "[...] the result of the listening test using the stylization obtained by the DG approach (Directed Graph (Nygaard and Haugland, 1998)) turned out to be similar to that of the DP approach, although DP achieves the minimum MSE."

Since statistical closeness measures do not appear to represent what happens on a perceptual level, Observation 2 gives credit to the choice made in D'Alessandro and Mertens (1995); Mertens (2004) to employ a tonal perception model to directly take into account psycho-acoustical phenomena in the pitch stylization task. This indicates to us the need to redefine the $q(S)$ function so that quality evaluation would be based on a tonal perception account rather than on the statistical closeness of S to P.

Observation 3. *Not all areas of the pitch curve should be stylized with the same degree of accuracy.*

The chi squared test on the subjective evaluation, together with the ANOVA test on Pps statistics regarding the OpSNoProm/OpSProm pair, shows that by exploiting prominence annotation, the stylization quality is not damaged from a perceptual point of view and uses significantly fewer points than the basic approach. This result indicates that saliency is an important feature to take into account when producing stylizations: by allowing the use of fewer points without influencing how the contour is perceived, the stylized curve complies with the requirements set by the definition and, consequently, better captures important pitch movements. The task of automatically identify salient areas, however, is not straightforward both from a theoretical and from a practical point of view. As a consequence of these considerations, the following observation is introduced:

Observation 4. *It is safer to employ acoustic parameters related to syllabification and prominence detection into an integrated control system for the stylization algorithm rather than introducing a serialized process performing segmentation and prominence annotation before stylization.*

In the preliminar tests, a manual prominence annotation performed by an expert linguist was used to identify salient areas. To make the OpS approach independent from manual annotation, the obvious approach would have been to introduce an automatic syllabification and prominence annotation step, as suggested in Origlia et al. (2011). The idea would have been of using the method presented in Petrillo and Cutugno (2003) to perform syllabification and combine it with the one presented in Ludusan et al. (2011) to perform prominence annotation. This approach, however, was discarded because it is not possible to obtain an automatic syllabification that always matches the manual one. Although past approaches like D'Alessandro and Mertens (1995); Mertens (2004), employed the concept of *phonetic syllable* to produce stylizations on the basis of the underlying syllabic structure, it is not possible to follow this approach because we also need to identify prominent syllables, introducing the problem of automatic prominence annotation: past approaches regarding automatic prominence annotation (Silipo and Greenberg, 2000; Tamburini, 2006; Abete et al., 2010; Avanzi et al., 2010; Ludusan et al., 2011) assume a manual syllabic segmentation to perform acoustic analysis leading to prominence detection. It would not be procedurally correct to apply an approach designed to be used on a manual syllabification with an automatic segmentation into *phonetic syllables*. Moreover, the definition of prominence itself is debated: some (Silipo and Greenberg, 2000; Avanzi et al., 2010; Ludusan et al., 2011) use a binary definition of prominence while others (Vanderslice and Ladefoged, 1972; Eriksson et al., 2002; Jensen, 2003; Tamburini, 2006) prefer to employ more levels. It is interesting, however, to note that in the scientific community there is much more consensus regarding the acoustic correlates of prominence: typically, energy and duration of the syllable nucleus together with synchronized pitch movements are taken into account to derive a prominence function in which local maxima correspond to prominent syllables.

In House (1990) the Spectral Constraint Hypothesis (SCH), which states that *"As the complexity of the signal increases, pitch sensitivity decreases"*, was introduced: tonal movement perception capability was described as inversely proportional to the amount of change in energy and spectral information independently by the type of change. In House (1995), however, results showing that prepausal tonal movements were perceived

with increased accuracy by human listeners were presented. When describing the effects of this finding, the author indicated that (House, 1995, p. 952)

> "[...] greater precision is necessary in modeling prepausal boundary tones for speech synthesis and automatic stylization of intonation than is necessary for phrase internal contours."

Following this finding, the author updated the SCH to take into account the syllabic structure by stating that (House, 1996, p. 2051)

> "[...] The area of maximum new spectral and intensity change occurring typically between syllable onset and syllable nucleus appears to be a crucial point for the timing of tonal movement. Movement through this area will be recoded as tonal levels as indicated in the earlier model. However, movement through the beginning of the syllable coda can be perceived as movement per se and described using movement features."

By phonetically reinterpreting the phonological framework used by House, it can be hypothesized that tone perception is related to the synchronized energy behavior both in terms of movement type (rising/falling) and in terms of rate of change (slopeness). There also seems to exist an extended superimposition between the effects described in House (1990, 1995, 1996), the widely used parameters for automatic prominence annotation and a syllable-like segmentation approach based on energy peaks and valleys. Because of this, an automatic pitch stylization algorithm can be built in such a way that every parameter is taken into account at the same time during the process without having to introduce pitfalls and further discretization of the speech signal based on a phonological account. The analysis of the interactions between pitch movements and energy glides should, of course, be at the basis of such an approach.

Observation 5. *Thresholds dealing with F_0 glissando perception should not be absolute but rather modulated by the interaction of pitch movements with co-occurring energy movements.*

There has been an extensive investigation, in the past, regarding the glissando threshold (Sergeant and Harris, 1962; Klatt, 1973; Schouten, 1985; t'Hart et al., 1990), the limit in the rate of change of a pitch segment over which a gradually changing tone is perceived instead of a static one. The differential glissando threshold, the limit in the difference between the rate of change of two segments over which they are perceived as two different glissandos, was less studied but it was taken into account in D'Alessandro and Mertens (1995). This parameter will be taken into account in the presented pitch stylization method as well.

Regarding the glissando threshold, the research effort concentrated on looking for an analytic expression that could tell if a pitch movement would be perceived as a static tone rather than a glissando. In t'Hart et al. (1990) such a expression was derived by comparing the different results that were obtained in previous works and it was defined as

$$g_{thr} = .16/T^2 \tag{2.9}$$

where T is the time interval in which the movement is realized. In Mertens (2004), however, it was found that, when comparing an automatic prosodic transcription, the *Prosogram*, with one provided by humans, the best result could be obtained by doubling the constant value at the numerator of Equation 2.9 and using it to produce a pitch stylization by the method presented in (D'Alessandro and Mertens, 1995) before performing the automatic transcription step. The way in which the formulas in t'Hart et al. (1990) and D'Alessandro and Mertens (1995) are constructed implicitly assumes that the threshold they define is intended to be absolute. As a matter of fact, in (t'Hart et al., 1990, p.33) Equation 2.9 was named *absolute threshold of pitch change*. In the same work, the authors had the intent of finding a glissando threshold that was as compatible as possible with the ones proposed in the previous studies they took into account. Even though this proposal was reasonably near to most of the examined studies, the authors also observed very high differences between their threshold and the ones presented in (Sergeant and Harris, 1962; Pollack, 1968; Rossi, 1971).

In this work, I adopt a different approach with regard to the glissando threshold and take into account a number of results about the effect the energy contour has on glissando perception (Zwicker, 1962; Maiwald, 1967; Feth, 1972; Rossi, 1972) to produce our stylization. Since energy movements can modify the way pitch glissandos are perceived, they must be taken into account when trying to algorithmically predict how a certain movement will be perceived by human listeners. Under this assumption, a mathematical model taking into account the psychoacoustical effect resulting from the interaction of energy and pitch movements should dynamically adjust the thresholds given the specific situation it is representing.

2.4.2 The tonal perception model

First of all, the rate of change of the segment $[s_x, s_y]$ is defined as follows:

$$V([s_x, s_y]) = \frac{|v_{s_y} - v_{s_x}|}{t_{s_y} - t_{s_x}} \tag{2.10}$$

The difference between the stylized segment and the original one can be evaluated by considering the difference between the two slopes.

When discussing Observation 2, I underlined the need of employing measures that would take into account psychoacoustical phenomena rather than statistics about curve closeness. In Observation 3 I also described how an efficient quality measure should not evaluate every part of the curve in a uniform way and, in Observation 4, I described the advantages of an integrated phonetic model with respect to a serial phonological model. Lastly, in Observation 5 I noted that, to correctly employ the psychoacoustical effects given by the interaction between energy and pitch, the glissando threshold and the differential glissando threshold should be automatically adjusted with respect to the specific situation. We now construct our quality function on these premises by using a number of accessory functions dedicated to the analysis of different aspects of pitch stylization based on a psychoacoustical background. First of all, we introduce a Γ_g function which evaluates the likelihood of a pitch segment to be heard as a glissando.

$$\Gamma_g([s_1, s_2]) = \begin{cases} 1 & if\ V([s_1, s_2]) > \frac{0.32}{T^2} \\ \\ \left(\frac{V([s_1,s_2])T^2}{0.32}\right)^{\gamma} & otherwise \end{cases} \tag{2.11}$$

This function is based on empirical proof presented in (Mertens, 2004) where 0.32 is the best value to use as numerator in Equation 2.9. However, we will not use this value to introduce an abrupt separation between movements that are perceived as glissandos and movements that are perceived as static tones. We start by assuming that if the rate of change of the pitch segment exceeds Mertens' threshold it will be perceived as a glissando and then, should this not be the case, we compute the likelihood of the pitch segment to be perceived as a glissando as the ratio between the actual rate of change and the one established in Mertens (2004). Another advantage of this method is that the original threshold established by t'Hart et al. (1990) is not simply ignored but it is rather assigned a glissando likelihood value of 0.5. In Equation 2.11 we also model the effect of energy movements on pitch perception by means of a gamma correction. The γ value which is used as exponent in the second part of equation 2.11 is evaluated as follows

$$\gamma = \begin{cases} \frac{100+mean(E')+1}{100} & if\ mean(E') \leq 0 \\ \\ \frac{100}{100-mean(E')+1} & otherwise \end{cases} \tag{2.12}$$

where E is the energy profile. In this equation, the value that will be put as exponent in the second part of Equation 2.11 depends both on the direction of the local energy profile and on its slopeness. The effect of the gamma correction is to dampen the modeled glissando perception capability proportionally to the slopeness of the energy profile if the pitch movement is aligned with an ascending energy movement. In the area comprising the phonetic syllables nuclei, where spectral stability is expected ($E' \approx 0$), there will be little or no gamma correction on the perceived pitch estimated in Equation 2.11 while pitch movements aligned with falling energy glides will be more likely to be heard as glissandos proportionally to the slopeness of the energy curve. We can describe the gamma correction

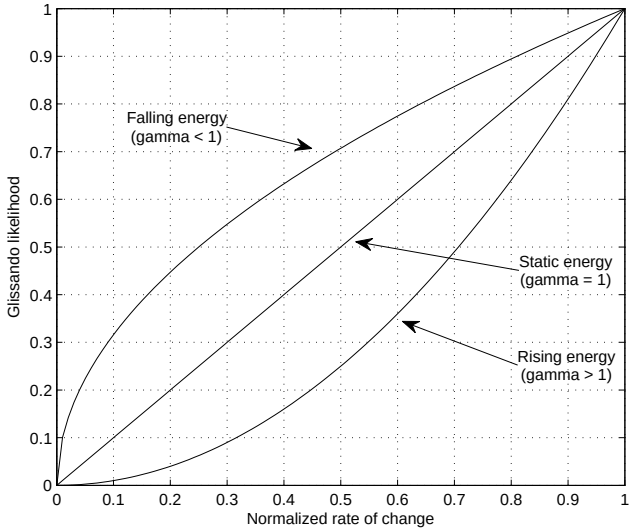


Figure 2.11: Glissando likelihood values are computed on the basis of energy movements in terms of gamma correction. In the figure glissando likelihood value transformations for glissandos not exceeding Mertens' threshold are reported.

effect as a controller for a dynamic gradient between the value which is assigned to flat pitch, which is 0, and the one assigned to movements reaching Mertens' threshold, which is 1. In conditions of spectral stability, the gradient is linear and the middle value corresponds to t'Hart's threshold. When spectral conditions change causing rising or falling energy movements, the middle value is reached respectively at higher or lower rates of change. This is summarized in Figure 2.11.

It should be noted that the effect of energy movements alone, while being consistent with the formulation of the SCH based on syllabic subparts, cannot account for the full set of changes indicated by the original formulation of the theory. Specifically, while the energy profile can detect changes in the amount of energy found in the spectrum, it cannot detect changes in the energy distribution among the frequencies. Adding to the model the capability of detecting and reacting to this kind of changes will be matter of future works.

For the sake of simplicity and in absence of contrary evidence, I assume that the sharpening effect in glissando perception when falling energy movements are present is symmetrical to the one found in correspondence of rising ones. I also assume that the maximum

energy difference we can find on an energy movement is 100 db. This value seems to be reasonable given the recordings found in the corpora in my availability.

We now define the difference between the glissando likelihood of two vectors $[s_i, s_{i+1}]$ and $[\bar{s}_i, \bar{s}_{i+1}]$ as follows

$$D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}]) = D_{acc}([s_i, s_{i+1}]) + \Gamma_g([s_i, s_{i+1}]) - \Gamma_g([\bar{s}_i, \bar{s}_{i+1}]) \tag{2.13}$$

where $D_{acc}([s_i, s_{i+1}])$ is the accumulated distance of S from P in the time interval $[t_i, t_{i+1}]$ after the preceding stylization steps. Let us clarify this with an example: suppose we want to stylize the curve $[s_1, s_2, s_3]$ and we find that removing s_2 is a good move. After merging $[s_1, s_2]$ and $[s_2, s_3]$ we obtain $[\bar{s}_1, \bar{s}_3]$ which can be also viewed as $[\bar{s}_1, \bar{s}_2, \bar{s}_3]$ where $\bar{s}_1$ and $\bar{s}_3$ are control points (they are equal to s_1 and s_3) and $\bar{s}_2$ is the result of the linear interpolation between the two in the time instant t_2. We can therefore define $D_{acc}([\bar{s}_1, \bar{s}_3])$ as the mean glissando likelihood difference of the segments in $[\bar{s}_1, \bar{s}_2, \bar{s}_3]$ from the ones in $[s_1, s_2, s_3]$. This method allows to keep track of the modifications made during the preceding steps and to correctly evaluate the impact of the new ones. Obviously, when $s_i = p_i$ and $s_{i+1} = p_{i+1}$, $D_{acc}([s_i, s_{i+1}]) = 0$ holds.

It is now possible to evaluate how likely it is that the stylization process has introduced a glissando in the $\bar{S}$ curve where a static tone was perceived in the S curve and vice-versa by defining a glissando quality evaluation function $q_g(S, \bar{S})$. Since the maximum value that the function $D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}])$ can assume is 1 by construction, it is straightforward to define the quality of a single segment of the $\bar{S}$ curve with respect to glissandos and static tones as $1 - |D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}])|$. By evaluating this formula on each segment in the S curve and weighting the results for the time fraction that the segment stylizes, we obtain

$$q_g(S, \bar{S}) = \sum_{i=1}^{n-1} \left((1 - |D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}])|) \frac{t_{s_{i+1}} - t_{s_i}}{t_{s_n} - t_{s_1}} \right) \tag{2.14}$$

As in the preceding example, I assume that if s_k is the point in S which is candidate for removal, $\bar{s}_k$ is the result of the linear interpolation of the points s_{k-1} and s_{k+1} in the time instant t_k.

The notion of differential glissando is now used to deal with the problem of checking that perceived glissandos are not being altered by the stylization process in a perceivable way. In fact, while Equation 2.14 can detect the erroneous introduction or removal of glissandos, it does not tell anything about the difference between two glissandos, even if their rates of change are not concordant in sign. We define the Γ_d function, which evaluates glissando similarity, as follows:

$$\Gamma_d([s_1, s_2], [\bar{s}_1, \bar{s}_2]) =$$

$$\begin{cases} \left(\frac{min(V([s_1,s_2]),V([\bar{s}_1,\bar{s}_2]))}{max(V([s_1,s_2]),V([\bar{s}_1,\bar{s}_2]))}\right)^{\frac{1}{\gamma}} & if\ concordant(V([s_1, s_2]), V([\bar{s}_1, \bar{s}_2])) \\ 0 & otherwise \end{cases} \tag{2.15}$$

In Equation 2.15, the value 0 is given to vectors which are not concordant in sign. If they are, the ratio between the two is considered, as it was in t'Hart et al. (1990), and the final score is modulated by the energy in the same way it is in Equation 2.11. In this case we use the inverse value of γ because while a falling energy movement, bearing sharpening effect, increases the glissando perception likelihood, evaluated by Equation 2.11, the same effect lowers the glissando similarity likelihood, which is estimated by Equation 2.15, due to the increased sensitivity we intend to model. For rising energy movements the inverse way of reasoning holds.

We can now define the likelihood of the $\bar{S}$ curve to introduce perceivable differences in glissandos with a $q_d(S, \bar{S})$ function

$$q_d(S, \bar{S}) = \sum_{i=1}^{n-1} \left(\Gamma_d([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}])\Gamma_g([s_i, s_{i+1}])\frac{t_{s_{i+1}} - t_{s_i}}{t_{s_n} - t_{s_1}}\right) \tag{2.16}$$

The value obtained by applying $\Gamma_d([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}])$ is weighted not only by the time fraction of the whole curve that the segment is stylizing but also by the value of $\Gamma_g([s_i, s_{i+1}])$.

This way Γ_d contributes to the quality evaluation of the $\bar{S}$ curve segments proportionally to the likelihood of the corresponding segments in S to be heard as glissandos.

We can now combine the $q_g(S, \bar{S})$ function and the $q_d(S, \bar{S})$ into a new quality function by taking their weighted mean

$$q(S, \bar{S}) = \frac{q_g(S, \bar{S}) + q_d(S, \bar{S})}{1 + \sum_{i=1}^{n-1} \left(\Gamma_g([s_i, s_{i+1}]) \frac{t_{s_{i+1}} - t_{s_i}}{t_{s_n} - t_{s_1}} \right)} \tag{2.17}$$

The $q(S, \bar{S})$ function represents a quality evaluation of the stylized curve with respect to the findings regarding tonal perception we took into account. This function can be introduced in the original framework of the OpS algorithm with little modifications that I describe in the next section.

2.4.3 Segmentation strategy

In the previous approach, the algorithm systematically divided in two equal parts each subcurve during the segmentation step. After introducing the new quality function, however, continuous rise/fall movements were not stylized well by the algorithm either by misaligning the peak, damaging quality, or by describing the movement with a plateau, damaging economicity. This was caused by the fact that the new quality function, being more flexible than the one based on NRMSE, sometimes was also less strict than needed when stylizing the small portion at the peak of the movement at the early steps of the algorithm. The problem was addressed by introducing the following rule in the segmentation strategy: if the curve contains a local maximum, the curve is splitted in the point corresponding to it, otherwise the curve is split into two equal parts. The new rule implicitly assigns more importance to local maxima: being evaluated later in the backtracking process, they are never considered in a limited context, which was the situation that caused the problem. This modification does not alter the *divide et impera* setup of the algorithm because the approach does not impose a specific segmentation strategy. Considering local minima as splitting points did not seem to introduce any improvement in the first set of

experiments and were therefore not considered for the results presented in Section 2.4.6. The modification is shown in Algorithm 2.

2.4.4 Cost function

The tonal perception model described in Section 2.4 is designed to dynamically adjust the quality evaluation of the curve given the pitch movement and its interaction with the energy profile. This method has a more solid foundation than the old, empirical, one based on the computation of the local variability index β which was used to set the balance of the harmonic mean in Equation 2.5. We can therefore weight equally the quality and the cost measure in Equation 2.5 by setting a constant value $\beta = 1$. Also, since we do not rely on a binary prominent/non-prominent manual annotation but on a dynamic qualitative model, we do not need anymore to alter the steepness of the sigmoid cost function in different situations.

Another modification we introduced addressed a problem of the cost function related to the fact that it was not making any distinction between long and short curves: cost differences were much more evident in short voiced segments than they were in long ones, causing the cost function to be too rapidly dominated by the quality function. To address this, we imposed a penalization factor to the value of the ratio between the original number of points and the one proposed by the stylization based on the portion of the entire curve that the subcurve is stylizing. Since in the original cost function high values for the ratio resulted in worse cost evaluation, the penalization factor is designed to give a slightly higher value to the ratio depending on the length of the voiced segment. To do this we compute an α value to be used as exponent for the ratio in the original cost function

$$\alpha = ((2 * (t_i - t_k))/(5 * (t_n - t_1))) + 0.4 \tag{2.18}$$

where t_i and t_k are, respectively, the end and start time of the subcurve and t_n and t_1 are the end and start time of the whole curve in the voiced segment. The formula constrains the penalization factor α in the interval [0.4,0.8]. This value was tuned on a

development set comprising 20 speech audio files that were not part of the corpora used for tests. The α value was chosen by empirically checking the balance between the quality of the resynthesis and the number of points used on the development set. Introducing α in the cost function, we obtain

$$c(S, \bar{S}) = 1 - \left(\frac{1}{1 + exp(\frac{-(x^{\alpha} - 0.5)}{0.1})} \right) \tag{2.19}$$

where $x = \frac{|\bar{p}|}{|p|}$.

2.4.5 Testing methodology

Because of the different approach, the testing procedure has to be adjusted. The main cause of incompatibility between the old testing methodology and the new approach lies in the fact that the stylized curve depends not only on how pitch behaves but also on how energy movements are synchronized with them. While this has no effect on the objective test, it is necessary to preserve the energy profile in the subjective test in order to correctly check the decisions made by the new algorithm and, because of this, it is not possible to use the *humming track* anymore. Utterances in the subjective test corpus were therefore resynthesized by using the PSOLA algorithm implemented in PRAAT and, to keep the subjects focused on pitch differences, they were explicitly asked not to care about what was actually being said but to look for differences in intonation.

This different setup also allowed us to align ourselves with the original test employed in t'Hart et al. (1990) in the sense that we could generate the stimuli in group C (the artificially altered ones) following the directions of the work presented by t'Hart: these stimuli were generated by shifting the pitch curve forward by 50ms rather than 10% up. This had the effect of introducing intonational mismatches that were subtle and localized just like the kind of errors a stylization algorithm usually tends to commit. This was intended to induce test subjects to pay particular attention to small, localized differences. The listening test setup was the same one presented in Section 2.3.5.

Since there was no qualitative difference among the competing algorithms in Origlia

Algorithm 2 *The OpS algorithm*

procedure $merge(P, S_a, S_b)$

$k = length(S_a)$

$z = length(S_b)$

$f_{orig} = F([s_{a_1}, ..., s_{a_k}, s_{b_2}, ..., s_{b_z}])$

$f_{mod} = F([s_{a_1}, ..., s_{a_{k-1}}, s_{b_2}, ...s_{b_z}])$

if $f_{orig} <= f_{mod}$ **then**

return $[s_{a_1}, ..., s_{a_{k-1}}, s_{b_2}, ...s_{b_z}]$

return $[s_{a_1}, ..., s_{a_k}, s_{b_2}, ..., s_{b_z}]$

end procedure

procedure $optimize(P, p_{start}, p_{end})$

if $p_{end} - p_{start} < 2$ **then**

return $P[p_{start} : p_{end}]$

if $\exists k | (p_{start} < p_k < p_{end}) \wedge (v_k > v_{k-1}) \wedge (v_k > v_{k+1})$ **then**

$p_{mid} = p_k$

else

$p_{mid} = floor(length(P[p_{start} : p_{end}])/2)$

$S_a = optimize(P, p_{start}, p_{mid})$

$S_b = optimize(P, p_{mid}, p_{end})$

$S = merge(P[p_{start} : p_{end}], S_a, S_b)$

return S

end procedure

procedure $main(P)$

for all voiced segments P_v **do**

p_{start} = Get the P_v start point index

p_{end} = Get the P_v end point index

$S_v = optimize(P, p_{start}, p_{end})$

end procedure

et al. (2011), it is necessary to compare the new OpS algorithm only against the old version employing manual prominence annotation. This way, the subjective test became much less tiring for the subjects.

Regarding the subdivision of the stimuli among the three groups, group A was differentiated between the two tests by swapping 5 stimuli with group C. This way, group A was composed both of shared stimuli and independent stimuli in a balanced way and, since group C was completely different, the test subjects could not become acquainted with the intentionally altered stimuli. The two test runs were administered to the subjects in different order to avoid the effect of overtraining and the stimuli presentation order was randomized during each test.

The number of subjects recruited for this test was 16 (8 males and 8 females). Five of them reported to have received some kind of musical training. By considering the performance obtained on the control group, two kind of statistics were computed: first I consider the whole group of human judges and then I consider only the subjects that performed sufficiently well on the control group by correctly evaluating at least 70% of the control stimuli.

Regarding the objective test, since the goal is to evaluate if the new approach introduces an improvement with respect to OpSNoProm and how different it is with respect to OpSProm, it is again not necessary to compare the new OpS algorithm with MOMEL and DP.

2.4.6 Results

First of all, it is evaluated how good the recruited human judges were in discriminating the stimuli in groups B and C. Among the 16 subjects, 5 were considered non-discriminative because they did not distinguish at least 70% of the control stimuli. As shown in Table 2.3, the performance of the discriminative group is higher: the mean values, which should ideally be 50%, show a better recognition capability with respect to the one of the full group. Table 2.3 also shows that, even among discriminative subjects, there is a certain

bias towards the *Equal* response. This is to be expected since the majority of the presented stimuli are intended to be labeled as such: they either are part of group B or they are generated by an algorithm designed not to introduce differences. All the subjects that received musical training were retained in the discriminative group.

Table 2.3: Control group statistics.

	Discriminative		Full group	
	Equal (%)	Different(%)	Equal (%)	Different(%)
Group B	90	10	87.5	12.5
Group C	27	73	43.13	56.88
Mean	58.5	41.5	65.31	34.69

Results of the subjective test are presented in Table 2.4. The chi-squared test did not show any significant difference in quality between OpSProm and the new approach. Given the design of the experiment and the results obtained both on the discriminative and on the full group of human judges, it is reasonable to assume that the qualitative performance of the two algorithms is the same.

Table 2.4: Subjective test results for both discriminative and discriminative + non-discriminative subjects. Percentages refer to the number of times a resynthesized utterance has been judged to be equal to the original one by the human listeners. Differences were never statistically significant by a chi-squared test ($p > 0.1$)

	OpS (%)	OpSProm (%)
Discriminative	78	83
Full group	79.38	82

Since we can assume that no perceivable difference exists between the curves produced by the new approach and the ones produced by the reference one, we concentrate on evaluating the difference in the number of points used. A repeated measures one-way ANOVA test on the number of points used in each file confirmed that it was possible to proceed with the paired t-tests ($p < 0.001$). The difference between the number of points

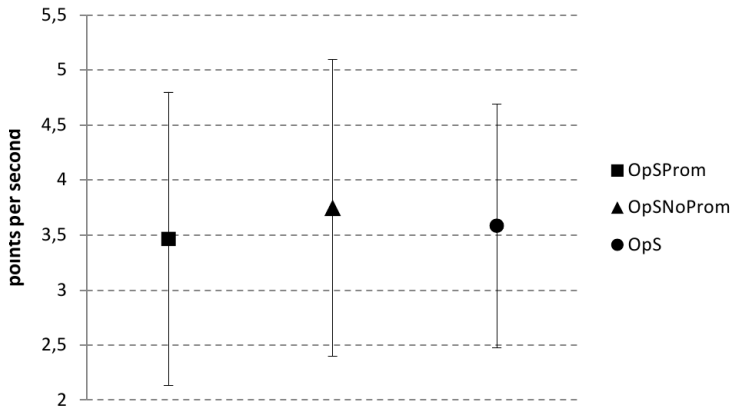


Figure 2.12: Mean values and standard deviations for the number of points per second employed by the different algorithms.

used by the new OpS algorithm has been found to be very significant with respect with the OpSNoProm algorithm ($p < 0.001$) while the difference with the OpSProm variant has not been found to be significant ($p > 0.01$). P-values were corrected using the Holm method. In Table 2.5 the results both in terms of points per second and in terms of the total number of points used to stylize the pitch curves of the test corpus are shown. These results were published in Origlia et al. (2013)

Table 2.5: Objective test results.

	OpS	OpS/Prom	OpS/NoProm
Points per second	3.59	3.47	3.75
Total points	4118	4009	4350

Curve economicity is particularly important to statistically investigate the macro-prosodic component of speech: by removing useless points, noise in the data is reduced and, if a point-based labeling system like INTSINT (Campione et al., 2000) is used, the number of labels needed to describe a movement is lower, thus simplifying analysis. Machine learning algorithms aimed at extracting information from prosody can also take advantage from this de-noised version of the pitch curve. The need for an automatic method capable to provide a perceptual account of intonational profiles has been underlined in recent years

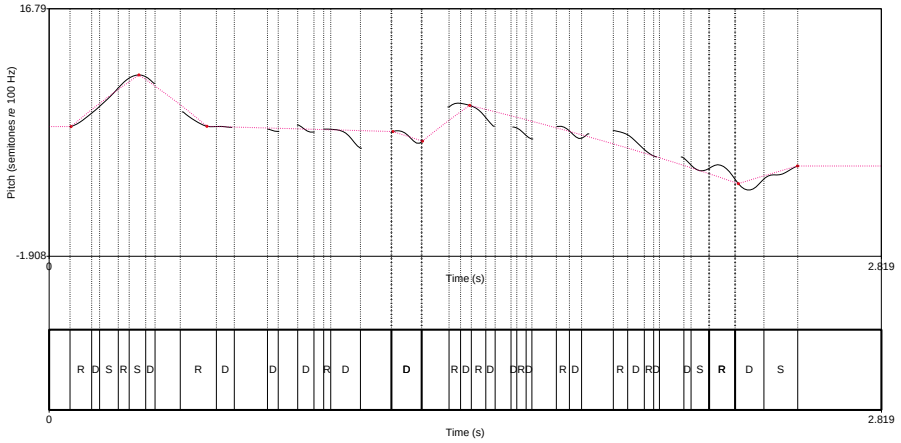


Figure 2.13: A stylization example. The pitch curve and the stylization proposed by the OpS algorithm are shown along with the energy profile of the utterance. A manual annotation of the energy profile is also shown: R indicates a rising energy movement, D indicates a descending one and S indicates an area of spectral stability.

by the applications that the *Prosogram* has found in different research areas (Patel, 2005; Ioannou et al., 2006; Caridakis et al., 2006; Avanzi et al., 2008).

The impact in basic research on prosody is highly significant too. A perceptually reliable stylization of F_0 constitutes a solid base on which to search for basic intonational units of natural languages. This is essentially the path followed by the IPO research group to find *standardized pitch movements* for different languages (t'Hart et al., 1990). Better automatic stylization techniques allow to continue on this path in a more reliable, consistent and fast way.

2.5 A simplified model: the SOpS algorithm

The results of the perceptual tests reported in the previous section, in which naive listeners were recruited, indicate that the stylization proposals of the OpS algorithm performed, in terms of quality, in a similar way with respect to other approaches. The OpS algorithm has the advantage of being parameter independent and it is able to use less points by explicitly taking into account a cost measure during computation. In Origlia and Alfano (2012),

we included the OpS algorithm in the Prosomarker tool (see Appendix A for details): an instrument designed to give a perceptual account of the pitch curves and to describe the synchronization of the pitch targets with automatically detected segmental events (syllable boundaries and nuclei). While using this tool to describe simple intonation phenomena, it was possible to trace a number of recurring situations in which the OpS algorithm was not able to capture specific classes of details from the curve that appeared to be critical to an expert linguists' ear.

2.5.1 Observations

Observation 6. *Considering local minima as equally important than non-maxima points did not make a difference for naive listeners but it introduced errors detected by expert listeners*

During the development of the OpS algorithm, it was found that giving priority to local minima if no local maxima can be found in the the curve during the splitting phase did not seem to introduce improvements. Not having this rule introduced the possibility that a local minimum was evaluated very early during backtracking. This implicitly assigns less importance to the point because the impact of its removal is evaluated on a limited portion of the curve. For the experts evaluating the quality of the OpS curves for their work this made a difference as they were able to detect small discrepancies both in timing and in tonal level of lowering targets in the resynthesis with respect to the original utterance.

Observation 7. *The quality measure dominating the cost measure in long pitch segments is not completely addressed by the introduction of the α parameter.*

Continuous pitch segments longer than the ones we tuned α on were found in other corpora: in these segments the effect was strong enough to make the α weighting useless. The presence of the α parameter is also less motivated from a theoretical point of view than the rest of the model, thus making the framework less reliable than intended.

Observation 8. *When local maxima split the curve in two subcurves that are very unbalanced in length, the algorithm was unable to adequately protect the smaller part of the curve.*

This was caused by the weighting of each segment dependently of the fraction of time it stylized. The quality of the longer subcurve was considered more important than the quality of the shorter subcurve that, subsequently, was often overstylized.

2.5.2 The final model

I now present the updates to the OpS algorithm that were introduced to address the problems highlighted in the previous section. The final model is simplified with respect to the preceding version. For this reason, I will refer to the final version of the OpS algorithm as the Simplified Optimal Stylization (SOpS) algorithm.

To address the problem presented by Observation 6, the splitting rule giving priority to local minima if no local maxima can be found was reintroduced. By evaluating these points later during the backtracking phase, the SOpS algorithm is able to protect low targets better than the OpS algorithm. Problems related to Observations 7 and 8 were both caused by the measure we used to evaluate shared endpoints removal during backtracking. Specifically, having the whole subcurves influence the quality measure introduced the problems related to differences in the curves' length. However, the removal of the shared endpoint, while generically influencing the quality of the two curves' mergings, is more specifically related to the quality of the two neighbouring *segments*. Back to the preceding example, given the A and B curves, the removal of the shared point $a_n = b_1$ only influences the quality of the $[a_{n-1}, a_n]$ and $[b_1, b_2]$ segments. Therefore, having the quality evaluation of the curves $[a_1, a_{n-1}]$ and $[b_2, b_m]$ contributing to the evaluation introduces an identical factor on both sides of the comparison operator. Eliminating this factor makes the algorithm take into account only the neighbouring segments quality. By weighting equally these two segments, the effect of longer movements being considered more important than shorter ones is removed too. Equation 2.14 is reformulated as

$$q_g(S, \bar{S}) = (2 - |D([s_{i-1}, s_i], [\bar{s}_{i-1}, \bar{s}_i])| - |D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}])|)/2 \tag{2.20}$$

Also, by considering local minima earlier in the splitting phase of the *divide et impera* schema, the *midpoint split* rule is applied to segments that are either quasi-linear or parabolic. In the first case, small differences are introduced by removing points while, in the second case, the *midpoint split* rule rapidly produces quasi-linear segments. This way, early evaluated points are more concerned with small details mainly depending by energy and pitch interactions, while lately evaluated points are more related with the description of larger prosodic events. Because of this distinction, it is not necessary to retain the fine details produced by the early backtracking steps up to the points controlling medium/long range pitch movements. Since the changes introduced by removing these points become very evident by delaying their evaluation to the latest steps of the backtracking process, the influence of the fine details in late steps of the decision process is not relevant. We therefore modified Equation 2.13 so that it does not keep track anymore of the preceding stylization steps obtaining the new formulation

$$D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}]) = \Gamma_g([s_i, s_{i+1}]) - \Gamma_g([\bar{s}_i, \bar{s}_{i+1}]) \tag{2.21}$$

Concerning the cost measure, as the impact of the sigmoid transformation revealed itself to have a negative effect with respect to the evaluation of the quality/cost balance, SOpS considers the untransformed ratio represented as x in Equation 2.19 as cost measure.

2.5.3 Evaluation

To evaluate the SOpS algorithm, we are interested in checking that the changes that were introduced to improve the performance on specific details do have an impact on these details without altering the general performance obtained with the OpS algorithm. The TIMIT dataset was used to check that the number of points used and the visual differences between the two algorithms are not relevant while a case study will be presented to show

that, on specific details, the output of the SOpS algorithm is better than the one obtained with the OpS algorithm.

From the quantitative point of view, we considered the number of points used by the SOpS algorithm with respect to OpS. The SOpS algorithm, on the considered dataset, uses 3.46 points per second (Pps) while the OpS algorithm uses 3.59 Pps. Table 2.6 shows a summary of the cost test between OpS, SOpS and OpSProm.

Table 2.6: Cost test results.

	OpS	SOpS	OpSProm
Points per second	3.59	3.46	3.47
Total points	4118	4007	4009

A paired t-test indicated that the difference in Pps between OpS and SOpS is not statistically significant ($\rho > 0.01$). However, close inspection of the pitch curves where the OpS algorithm introduced more points than necessary showed that the SOpS algorithm does not suffer from this problem. The amount of reduction observed (0.13 Pps) and the actual ρ-value (0.012) are coherent with the goal of reducing the number of points used only in specific areas. The performance of the SOpS algorithm in terms of Pps is much more similar to the one we obtained with the OpSProm algorithm. A paired t-test between the Pps measures obtained by SOpS and OpSProm confirms this ($\rho > 0.9$) with greater certainty with respect to the result presented in the preceding section, where the difference between OpS and OpSProm, while still not significant, ($\rho > 0.01$), was to be taken carefully as the actual ρ-value was 0.0142.

From the qualitative point of view, a Wilcoxon test on the differences between curves generated by the two algorithms showed that the location shift is not significant ($\rho > 0.4$). The size of the considered dataset makes it safe to assume that no significant differences can be found between the curves proposed by the two algorithms on a large scale. This result confirms that the modifications introduced by the SOpS algorithm do not alter the stylized curve up to a statistically detectable degree. Close inspection of the cases on

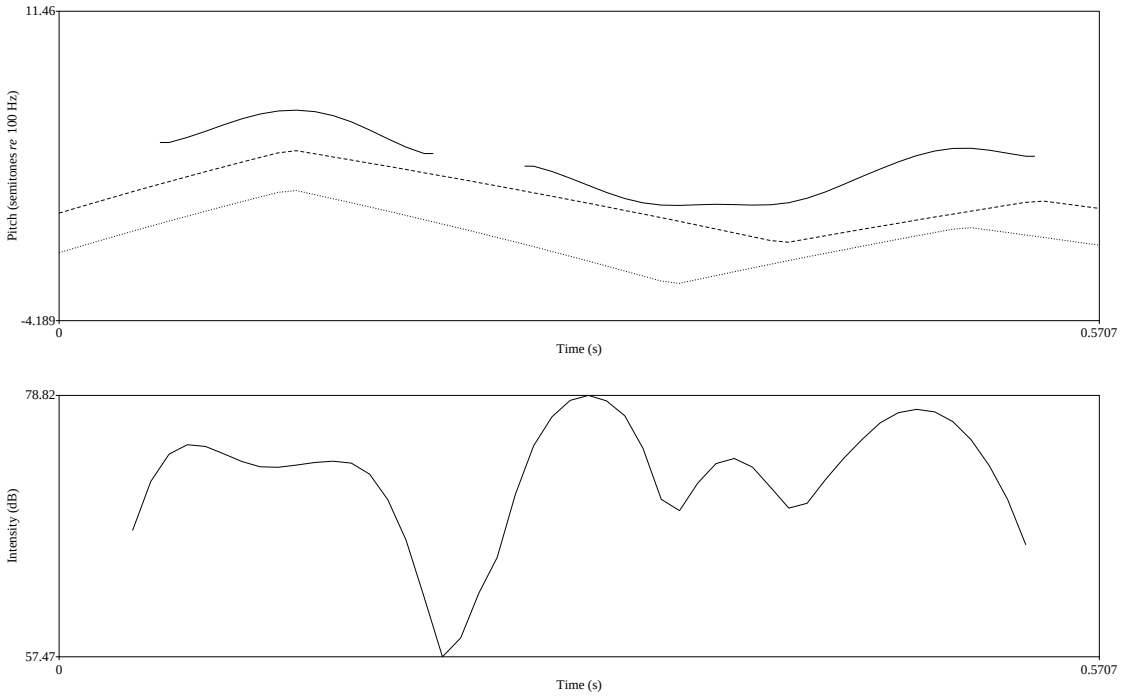


Figure 2.14: A pitch contour (solid line) along with the OpS stylization (dashed line) and the SOpS one (dotted line). The stylized curves are shifted by 2 Semitones each with respect to the original one for visualization purposes. Along with the pitch curve, the energy profile of the considered speech fragment is shown.

which the new model is intended to perform better, however, shows that the details the OpS algorithm was not able to retain are correctly modeled by the SOpS algorithm.

I will now present a case study to show the kind of modifications the SOpS algorithm introduces with respect to the OpS algorithm. In Figure 2.14, we show the detail of a pitch contour, the stylization proposed by the OpS algorithm (dashed line) and the alternative proposed by the SOpS algorithm (dotted line) along with the energy profile. While the two algorithms perform identically on the first movement, the final rise/fall sequence is described differently. Since the curve's portion after the peak is much shorter than the rest of the curve, protecting the final lowering movement was considered not valuable enough by the OpS algorithm. This decision is encouraged by the tonal perception model as the rising movement preceding the final fall is synchronized with a rising energy profile,

thus lowering the modeled glissando perception capability. The influence of sections that do not depend by the point being evaluated also plays a role, as discussed before. The SOpS algorithm, by considering only the neighbouring subcurves and by weighting them equally, is able to protect the final movement when evaluating the peak point, as expected because of the synchronized falling energy contour. The turning point before the rise is shifted 60ms earlier because of the segmentation strategy giving more importance to local minima. This improves the representation of the subcurve synchronized with the falling energy movement. The following pitch rise, synchronized with a rising energy contour, is more stylized than before, so no points are added. From perceptual inspection, this choice appears to improve the overall quality of the curve used in the example.

The magnitude of the changes the SOpS algorithm introduces with respect to the OpS curves are, in general, similar to the ones shown in the example. This explains why the similarity test based on statistical closeness is not able to detect a significant difference between the two algorithms. Being these changes important for an expert listener, however, we can confirm that statistical closeness measures are not good estimators of the general quality of a stylized curve.

2.6 Prominence detection

Linguistic research has concentrated for a long time on the investigation of syllabic prominence, the phenomenon by which some units are perceived to be salient with respect to the others. There is no consensus regarding the definition of syllabic prominence nor on the appropriate annotation methodology. A first definition was given in Bloomfield (1933):

> "[...]Stress - that is intensity or loudness - consists in greater amplitude of sound-waves, and is produced by means of more energetic movements, such as pumping more breath, bringing the vocal chords closer together for voicing, and using the muscles more vigorously for oral articulation"

This first definition concentrated on the energetic component of prominence, exluding any contribution from intonational strategies. This was rectified by Bolinger (1958) with the introduction of the *pitch accent*: an intonational marking realized on a particular unit that is therefore perceived as different from its surrounding ones. The most widely accepted definition of prominence is also a very prudent one, given by Terken (1991):

> "[...] Prominence is the property by which linguistic units are perceived as standing out from their environment."

This definition emphasizes the perceptual nature of prominence and it refers to a generic *environment* giving the prominent syllable a background with which it contrasts with. It is common, especially in automatic annotation studies, to use a binary notation to mark syllables (prominent/non-prominent). This type of annotation is generally preferred because it offers a simple method to evaluate the performance obtained by automatic approaches. Prominence detection has been the subject of a wide number of investigations in the past (Silipo and Greenberg, 1999; Tamburini, 2006; Abete et al., 2010; Avanzi et al., 2010; Ludusan et al., 2011). Automatic prominence detection systems are based mainly on rule-based approaches as machine learning techniques can make it difficult to understand how a certain performance was reached by the underlying statistical model. Although supervised approaches were used in this work, absolute performance was considered a secondary objective with respect to the possibility of using machine learning to collect data useful to improve current rule-based annotation systems based on a linguistic, as opposed to a statistic, background.

Conditional Random Fields (CRF) (Lafferty et al., 2001) are a class of discriminative models used for sequence segmentation and labeling which are designed to maximize the conditional probability of the labels given the sequence of observations. The use of CRFs is now well established as they have been successfully applied to a wide range of scientific fields, including natural language processing and speech analysis tasks. In the case of prominence annotation, it was shown that CRFs outperform HMMs in the task of predicting pitch accents at word level with a combination of acoustical and syntactic features

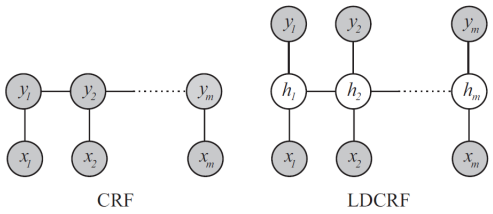


Figure 2.15: Graphical representation of a Conditional Random Field and a Latent-Dynamics Conditional Random Field.

(Gregory, 2004). CRFs were also used to investigate pitch accent detection along with the realization of givenness and focus at word level by employing lexical and acoustic features (Sridhar et al., 2008). Differently from these previous studies, in this work only acoustic features are used and the syllable level is taken into account.

There are three main ways in which this particular kind of sequence labeling models can be applied to the problem of automatic syllabic prominence annotation. **Structural differences analysis** among classifiers, paired with performance comparison, gives information regarding the interactions among the features. **Feature sets comparison** and **multiple contexts comparison** estimate the predictive power of the considered features and the amount of context that should be taken into account. In this work, these three different kinds of analysis are applied by employing different classifiers, feature sets and context extensions.

2.6.1 Latent-Dynamic Conditional Random Fields

CRFs are designed to capture inter-class relationships by maximizing the conditional probability of the sequence of labels from a sequence of observations. Given a set of weights estimated during training λ, the sequence of labels Y and the sequence of observations X, a Linear Chain Conditional Random Field estimates $P(Y|X)$ as follows

$$P(Y|X,\lambda) = \frac{1}{Z(X)} exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right) \tag{2.22}$$

where N is the number of observations, $f_k(y_t, y_{t-1}, \mathbf{x}_t)$ represents either a *state feature*

function or a *transition feature function* and $Z(X)$ is a normalization constant. State feature functions describe the relation between observation/label pairs while transition feature functions describe the relation between observations and transitions from one state to another. Since the definition of feature functions includes a vector of observations in the third term, the set of feature functions can be computed over an arbitarily extended context of surrounding observations W. CRFs are limited as they can model inter-class relationships but cannot model intra-class dynamics. Latent Dynamic Conditional Random Fields (LDCRF) (Morency et al., 2007) are an extension of CRFs designed to introduce hidden variables in the model, in order to capture both kinds of dynamics. Hidden states represent a sequence of unobserved variables H and are used to define the following latent conditional model:

$$P(Y|X,\lambda) = \sum_{H} P(Y|H,X,\lambda)P(H|X,\lambda) \tag{2.23}$$

The above model allows only disjoint sets of hidden states for each class label. Therefore, each label y_j has an associated set H_{y_j} of hidden states with $H_{y_i} \cap H_{y_j} = \emptyset$ for $i \neq j$, making it is possible to rewrite Equation 2.23 as:

$$P(Y|X,\lambda) = \sum_{h \in H_{y_j}} P(H|X,\lambda) \tag{2.24}$$

The conditional probability of the hidden states given the set of observations and weights can then be formulated as for the CRF model:

$$P(H|X,\lambda) = \frac{1}{Z(X)} exp\left(\sum_{k=1}^{K} \lambda_k f_k(h_t, h_{t-1}, \mathbf{x}_t)\right) \tag{2.25}$$

A graphical comparison between a CRF and an LDCRF is shown in Figure 2.15. In the LDCRF model there is no longer a direct connection between observations and labels due to the introduction of a layer of hidden variables. Since the labels are disconnected from the observations, they are considered to be conditionally independent, given the hidden states.

2.6.2 Feature sets

Features related to energy, segments durations and internal pitch movements for each syllable are usually employed in the automatic prominence annotation task and the particularly important role that syllable nuclei play in the detection of prominent syllables is widely recognized in the literature. In Silipo and Greenberg (1999), the mean amplitude inside the syllable nucleus ΔA and the nucleus length ΔT_n were combined into an evidence variable as follows:

$$Ev = \Delta A \Delta T_n \tag{2.26}$$

After computing an Ev value for each syllable, local maxima in the sequence of evidence variables were marked as prominent. Since, in the literature concerning automatic annotation of syllabic prominence, a great importance has been assigned to ΔA and ΔT_n, these two features are always included in the feature sets used in the experiments. Given the manual segmentation into syllables, nuclei onsets and offsets are estimated by taking the energy peak inside the syllable and computing the -3dB band.

Energy and duration do not account for prominences caused by pitch movements through the nucleus. In Avanzi et al. (2010), a syllable was automatically marked as prominent if a rising pitch movement exceeding a threshold was detected. In Abete et al. (2010), the same concept was implemented as an integration of the approach proposed in Silipo and Greenberg (1999) with a pitch movement dependent parameter. Equation 2.26 was then reformulated as

$$Ev = m \Delta A \Delta T_n \tag{2.27}$$

where m represented a heuristically computed penalization factor for syllables that do not exhibit a rising movement through the nucleus. The m parameter is included in the feature sets F3 and F4.

The previous two attempts to use pitch features in an automatic system for prominence annotation were based on heuristics and assumed that only rising pitch movements had an

Table 2.7: Feature sets composition

	ΔA	ΔT_n	m	Γ_{s_1,s_2}	ΔT_s	V/T_s
F3	✓	✓	✓			
F4	✓	✓	✓		✓	
F5	✓	✓		✓	✓	✓

effect on prominence perception. To check the informative content of the $\Gamma_{s_i,s_{i+1}}$ feature introduced in Equation 2.11, I substitute the heuristically computed m parameter with $\Gamma_{s_i,s_{i+1}}$ in the features set $F5$ together with the ratio between voiced time and total syllable time to give an account of the context in which the movement is realized, as it can influence perception.

The total syllable time ΔT_s is introduced in the features set F4 as it is common, in the literature, to find documentation regarding the importance of duration features in prosodic analysis. A detailed description of the composition of the three feature sets tested in this work is shown in Table 2.7.

2.6.3 Materials

For the experiments on automatic prominence annotation an Italian corpus containing read numbers and an English corpus containing read sentences were used. The Italian corpus consists of a subset of the SPEECON corpus (Siemund et al., 2000) that has been used to evaluate the system presented in Abete et al. (2010); Ludusan et al. (2011). The English corpus consists of a subset of the TIMIT corpus that has been used to evaluate the system presented in Tamburini (2006). Both subsets were manually segmented into syllables and annotated by an expert linguist using a binary notation for syllabic prominences. The SPEECON subset contains 288 utterances (15 minutes of speaking time) containing at least 5 syllables (mean: 15, total: 4265). The TIMIT subset contains 382 utterances (17 minutes of speaking time) containing at least 4 syllables (mean: 12.51, total: 4780).

Table 2.8: F-measures obtained on the SPEECON subsets. The best performance obtained with each features set by the two classifiers is marked in bold.

	CRF			LDCRF		
	F3	F4	F5	F3	F4	F5
W1	65.12	79.79	82.76	65.09	79.69	82.75
W2	67.10	82.08	84.30	75.10	84.66	86.32
W3	68.00	83.46	85.77	75.39	85.58	87.82
W4	68.27	84.05	**86.15**	76.12	**85.71**	87.66
W5	**68.94**	**84.07**	86.08	**76.55**	85.54	**87.85**

Table 2.9: F-measures obtained on the TIMIT subset. The best performance obtained with each features set by the two classifiers is marked in bold.

	CRF			LDCRF		
	F3	F4	F5	F3	F4	F5
W1	53.43	68.50	68.91	53.52	68.56	68.90
W2	60.47	71.71	70.96	72.03	**78.01**	77.24
W3	60.43	71.91	71.54	**72.52**	77.83	**77.52**
W4	61.46	72.07	72.03	72.12	77.86	77.26
W5	**61.76**	**72.36**	**72.48**	72.12	77.83	77.21

2.6.4 Results

Each classifier was tested on the SPEECON and on the TIMIT subset. For each features set, both classifiers were tested by varying the context extension for building the feature functions from a minimum of 1, which considers only the two neighboring syllables, to a maximum of 5. Performance is measured in terms of F-measure (Prominent class as TRUE) and the test protocol is 10-fold cross validation. The summary of the results obtained on the SPEECON subset is reported in Table 2.8 while results obtained on the TIMIT subset are detailed in Table 2.9.

Table 2.10: Statistical significance tests on the SPEECON corpus, for different pairs of values for the W parameter. Check marks indicate significant differences.

		Window length pairs					
		2/3	2/4	2/5	3/4	3/5	4/5
CRF	F3						
	F4	✓	✓	✓			
	F5	✓	✓	✓			
LDCRF	F3				✓	✓	
	F4				✓		
	F5	✓	✓	✓			

The statistical significance of the differences between the obtained performances was evaluated by means of a McNemar test. To evaluate the performance of the LDCRF with respect to the CRF, I compared, for each features set, the results obtained by the best CRF with the ones obtained by the best LDCRF. The differences were found to be statistically significant in all cases ($p < 0.01$).

To evaluate the performance difference obtained with the various feature sets, I compared the performance of the best LDCRF from each features set with the performance of the best LDCRFs from the other feature sets. While on the SPEECON subset the differences were found to be always significant ($p < 0.001$), on the TIMIT subset the differences $F3/F4$ and $F3/F5$ were statistically significant ($p < 0.001$) while the difference $F4/F5$ was not significant.

To evaluate the influence of the context extension, I compared, for each corpus, classifier and features set, the pairwise combinations of values of the W parameter. While on the TIMIT subset only comparisons involving $W = 1$ were found to be significant, tests on the SPEECON subset yielded a different situation, summarized in Table 2.10.

Both LDCRFs and CRFs applied to the SPEECON subset outperform the systems presented in Abete et al. (2010), in which 73.3% F-measure was reported, and in Ludusan

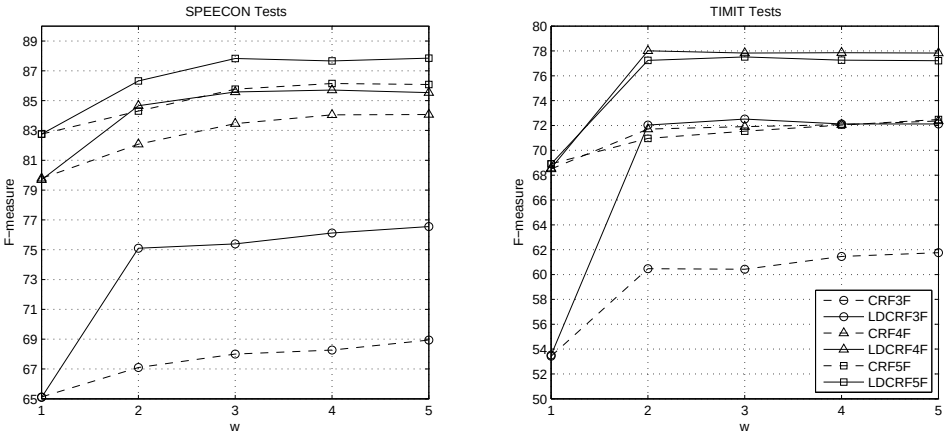


Figure 2.16: Summary of the obtained performances for each combination of classifier, features set and context extension on the two test corpora.

et al. (2011), where 75.1% F-measure was reported. Concerning the TIMIT corpus, in Tamburini (2006) an error rate of 18.64% was reported. If we take into account the performance of the LDCRF that obtained the best results on TIMIT ($F4/W2$), the error rate is 16.32%. A graphical summary of the presented tests is shown in Figure 2.16.

By varying the context extension, on the SPEECON subset significant differences among various tests can be found consistently up to a three syllables context. On the TIMIT subset a context window of two syllables seems to be sufficient to achieve maximum performance. This is in contrast with approaches used in earlier automatic prominence annotation (Silipo and Greenberg, 1999), but it is consistent with more recent findings regarding context extension (Avanzi et al., 2010).

By varying the composition of the features set, it was found that syllable length is a particularly important feature as the comparison between the best LDCRF using the $F3$ features set is always inferior to the performance obtained by the best LDCRF using the $F4$ features set in a statistically significant way. Since the $F5$ features set enabled the LDCRF to obtain better performance with respect to the $F4$ features set on the SPEECON subset only, the combination of the $\Gamma_{s_i,s_{i+1}}$ and V/T_s features contains at least the same amount of information as the m parameter, while having a better theoretical background.

2.7 Conclusions

I will now summarize the different elements that will constitute the basis of the prosodic representation used in the next Chapter.

Concerning syllabification, I substituted the phonological concept of syllable, typically used in prosodic studies, with its phonetical interpretation to obtain an automatic segmentation of the examined utterance. While the units constituting this segmentation only partially overlap with manually marked syllables, by using a phonetic template the system does not depend on language-specific syllabification rules, thus generalizing well. Also, as the *phonetic syllable* template has a very simple formulation, it does not impose a heavy computational load and it makes it easy to extract features because of its stability.

Concerning pitch stylization, by looking at the differences in terms of Pps among the considered algorithms, one could easily underestimate them, but the total number of points helps in evaluating the impact of the new approach and to understand why the statistical test detects a significant difference. For example, having OpS using 0.16 Pps less than OpSNoProm has the effect of removing 232 points without compromising perceptual equality while a difference of 0.28 Pps between OpSNoProm and OpSProm causes 341 points to be removed. If we interpret the stylization process as a filter for inaudible pitch movements, it is clear that the presented approach is able to improve the curve economicity without compromising perceptual equality.

As noted by t'Hart et al. (1990) and as shown in Section 2.3, visual similarity between the original curve and the stylized one can be misleading: in many cases the algorithm proposed a curve that looked over-stylized but was judged to be equal to the original one by the majority of the subjects participating in the perceptual test. In Figure 2.13 a pitch curve along with its stylization and a manual annotation of rising, falling and stable energy segments in voiced areas is shown. In the first highlighted segment, labeled D because of the descending energy glide, we can see that the algorithm chose to keep both endpoints of the voiced segment because it detected a significant energy drop along with the pitch

movement. On the contrary, in the second highlighted segment, labeled R because of the rising energy glide, the algorithm chose not to stylize the synchronized pitch movement. Although a number of visual differences can be noted in the stylized curve with respect to the original one, from the perceptual test I observed that 12 of the 16 judges labeled the proposed curve as equal to the original one. The same curve was judged to be equal to the original one by 9 of the 11 discriminative judges. This shows that, by removing the pitch movement aligned with a rising energy glide and protecting the pitch movement aligned with a dropping energy glide, the algorithm was able to mantain perceptual equality and to avoid using points stylizing a movement that seems not to have importance to the human ear.

An example of the potential of the automatic approach for linguistic research is Oliver (2005), in which a modified version of the MOMEL algorithm (Hirst and Espesser, 1993) is used to obtain a stylized F0 curve which is given as input to a clustering algorithm (Oliver, 2005, p. 161)

> "[...] to derive prototypical pitch contour types found in Polish, based on their acoustic characteristics."

On a similar line, Mertens (2006) matched default intonational profiles predicted by lexical and syntactic features with profiles actually produced by speakers of a corpus and automatically labelled by means of the *Prosogram*, in order to identify the marked movements of pitch which are not predictable from lexicon and syntax, thus conveying an independent meaning. Defining a set of descriptive intonational units in speech is a very important step both for research on prosody and speech technologies, however there is not much agreement on the nature and number of such units (Silverman et al., 1992; Campione et al., 2000; Taylor, 2000). Research in this field will continue in the coming years and will hopefully benefit from a stylized representation of the actually perceived pitch movements, which seems to constitute a better basis than simple F_0. Moreover, this kind of reliable pitch stylization allows to develop systems for automatic prosodic annotation, thus overcoming the limits of manual annotation, which is a time-consuming activity implying a

certain amount of subjectivity.

Concerning prominence detection, results offer insight on three different issues regarding prominence detection: model performance, context influence and feature sets. LDCRFs perform systematically better than CRFs. On the SPEECON subset, every LDCRF performs better than its direct CRF counterpart while on TIMIT this effect is even more evident as the lowest performance obtained by an LDCRF is similar to the best CRF performance. The main difference between the two classifiers lies in the presence, in the LDCRF model, of hidden states. This difference is critical as it allows the classifier to learn complex dynamics that are not explicitly described by the raw sequence of observations. In order to better understand these results two observations are important: (1) the advantage of LDCRFs is that they detect hidden dynamics inside a single class and (2) the binary annotation mainly produces sequences of non-prominent syllables separated by prominent syllables. A possible cause of LDCRFs outperforming CRFs in this task could, therefore, be that a hidden dynamic lies in the sequence of non-prominent syllables.

Chapter 3

Emotional speech

Human affect is a rich source of information, one of these consisting of audio signals. It is well known that affective information through audio signals is conveyed as a sum of explicit messages, consisting of semantic units, and of more implicit messages reflecting the way words are uttered, consisting of acoustic and prosodic cues. Affective information in everyday life as such a product may become very complex and variable. In this Chapter I will show how the syllable based speech analysis method presented in Chapter 2 performs on the dimensional and continuous emotional speech tracking task.

3.1 Non-verbal communication of emotions

In a specific study about the influence of nonverbal communication over human interaction it was shown (Graham et al., 1991, p. 59) that

> "[...] Nonverbal communication was important to all surveyed, and most respondents agreed that nonverbal communication would influence their interactions with people more than would verbal content."

In the same work, inconsistencies between verbal and nonverbal content were found to be severely disruptive when it comes to trust and mutual understanding. Being the

area of interest of the authors concerned with interactions in business organizations, the obtained results led to a very specific direction for managers to improve communication with employees (Graham et al., 1991, p. 58)

> "[...] Keeping in mind that such discrepancy can cause miscommunication, distrust, and frustration managers should become more cognizant of this problem and make real efforts to keep their verbal and nonverbal communication consistent with each other."

There are a number of issues in emotion research that still remain unsolved, as recently highlighted in Schuller et al. (2011). Focusing here on the task of automatically detecting emotional content from purely acoustic properties in the human voice, two points seem to be dominating the debate at present. The first point is represented by the need of a shared view of emotions in terms of the way they should be collected and modelled. The second one is related to the definition of what can be considered the smallest chunk of analysis to be referred to for evaluation procedures on emotional corpora (real or acted). This problem must take into account the needs of real-time systems needing to give an estimate of the emotional content of an utterance during its production rather than waiting for it to be completed.

The first point was discussed in Chapter 1. Going to the second point, different approaches have taken into account different chunks of speech for emotional speech analysis and classification (ranging from utterances to smaller units such as vowels and syllables, i.e., linguistic units, or other technical units such as frames or time slices (see Schuller et al. (2011)). Features extraction has been carried out on whole utterances (see Fragopanagos and Taylor (2005) for a review); on segments like words, *ememes*, considered the smallest possible meaningful emotional unit (Batliner et al., 2010); on syllables obtained through Forced Alignment (FA) or Automatic Speech Recognition ASR (Kao and Lee, 2006); on individual vowels investigating the effect emotion dimensions have on formants placement (Goudbeek et al., 2009) or accounting for the importance of formants in the distinction of emotions (Vlasenko et al., 2011; Gharavian et al., 2012), analyzing the voice quality char-

acteristics of stressed and unstressed vowels pronounced in a VCV setting (Drioli et al., 2003), exploring the correlation of a set of discrete emotions with a number of acoustic features extracted from sustained \\a\\ sounds (Patel et al., 2011); on distinct portions of speech using a Voice Activity Detection module to extract features only in active speech areas (Wu et al., 2009, 2010), thus working at utterance level. Related to this second point and comparing earlier with more recent works on automatic emotion recognition, there is a tendency to limit or reduce the number of features used for classification or for recognition. For example, in Vlasenko et al. (2011) the authors compare their recognition results to those presented in Schuller et al. (2008) stressing the fact that they used only a two class problem instead of a seven class problem, only one average F1 value instead of 39 MFCC, 7 indicative vowels instead of 41 phonemes and, finally, only one Gaussian for each phoneme model instead of 96. Such comparisons may look like a cutthroat competition, but to obtain real-time systems estimating the emotional content of an utterance during its production, an important step lies in reducing the amount of features to be analyzed, the computational difficulties implicit in each modelling attempt and possibly, as in the case of the present work, the amount of data to be analyzed. In this work, we will adopt an approach focusing on syllable-like segments as our smallest unit for features extraction. These particular units are automatically detected as shown in Section 2.2 and coincide with the concept of *phonetic syllables* by D'Alessandro and Mertens (1995):

> "[...] a continuous voiced segment of speech organized around one local loudness peak, and possibly preceeded and/or followed by voiceless segments"

However, while keeping the term *phonetic syllables*, I will follow a later definition, given by Roach (2000), that accounts for voiced consonants in a better way. In (Roach, 2000, p.70), in fact, these units are described as

> "[...] consisting of a centre which has little or no obstruction to airflow and which sounds comparatively loud; before and after that centre (...) there will be greater obstruction to airflow and/or less loud sound"

Since another possible issue in emotion classification is, as above mentioned, represented by features extraction techniques and types/number of features considered, it is necessary to pay particular attention to the features extraction technique and on the features used to perform automatic regression. In this work, I will present a features extraction technique different from the ones usually found in the literature in the sense that supra-segmental features extraction will be strictly connected to a preliminar segmental analysis, following the general indications coming from linguistic studies on prosody in which intonational phenomena are often described in terms of the synchronization between pitch movements and the occurrence of segmental units.

3.2 Features set

After segmenting the speech signal into *phonetic syllables* and stylizing the pitch contour by means of the SOpS algorithm (Section 2.5), features are extracted from each syllable nucleus at least 64ms long are retained. This value corresponds to the window length used to compute the Intensity profile inside the syllable nucleus at a higher resolution so that Shimmer can be evaluated. Concerning duration features, nucleus and syllable lengths are included together with the length of the syllable's head and coda. Of course, these values are 0 if the corresponding segment is not present. The mean pitch, energy, Harmonics-Noise Ratio (HNR) and Zero Crossing Rate (ZCR) are also extracted together with the mean value of the Teager Energy Operator (TEO). Shimmer is extracted as a measure of stability for the energetic content. As the window length used to compute Shimmer is 64ms long, only syllables having a nucleus equal or longer to this value were retained. From the stylized pitch contour, the mean value in semitones together with the glissando likelihood value of the segment crossing the energy peak, computed as shown in Equation 2.11, is included in the features vector of each syllable. Therefore, for each *phonetic syllable* associated with a nucleus represented by a voiced energy peak occurring at time t_k, if $[s_i, s_{i+1}]$ is the stylized segment such that $t_{s_i} < t_k < t_{s_{i+1}}$ holds, $\Gamma_g([s_i, s_{i+1}])$ is included in the features set of the considered syllable. This is intended to describe the

likelihood of a pitch movement realized in a syllable nucleus to be heard as a glissando.

The first 13 MFCC coefficients are extracted from syllable nuclei only at frame level. The MFCC vector is extracted by centering a 15ms window on the energy maximum associated with the syllable nucleus. This is because we can assume that energy distribution inside the nucleus is relatively stable across the frequencies.

Taking as reference the work presented in Wu et al. (2010), we consider to be *prosodic* all the features that do not represent energy distribution among the frequencies. That is, only duration, energetic and periodicity related features are considered to be prosodic. This can be considered very restrictive and possibly inaccurate as it leaves out of the picture voice quality that is considered a supra-segmental characteristic of speech (Scherer et al., 2003). We follow this approach to allow comparison between our results and the ones presented in Wu et al. (2010) for what it concerns the performance of *prosodic* features only. The representation of spectral information, in this work, is delegated to MFCC features to allow comparison between the performances of *prosodic* features combined with MFCCs, which in Wu et al. (2010) yields the best results on the VAM corpus.

To allow the use of SVMs to perform emotion regression, syllable-level features are collapsed into global statistics, similarly to what it was done in Wu et al. (2010) after extracting frame-level features. During speech production, however, not all segmental units have the same importance: prominent syllables, generally defined as *syllables standing out with respect to their context*, contain more reliable acoustic information than their non-prominent counterparts (Seppi et al., 2010). Since a number of works concerning the automatic detection of prominent syllables (Silipo and Greenberg, 1999; Tamburini and Wagner, 2007; Avanzi et al., 2010; Ludusan et al., 2011) highlighted the particular importance of the syllable's nucleus length, when extracting global statistics from syllable-level features we compute the weighted mean of all features by taking the normalized nuclei lengths as weights as shown in the following equation:

$$w_i = n_i / \sum_{j=1}^{N} n_j \tag{3.1}$$

where w_i is the weight of the features extracted from the i-th syllable, n_i is the length of the i-th nucleus and N is the number of syllables in the utterance. Since the weights sum to 1, mean values of each feature are obtained as follows:

$$\bar{\mu}_j = \sum_{i=1}^{N} f_{i_j} w_i \tag{3.2}$$

where $\bar{\mu}_j$ is the weighted mean of the j-th feature and f_{i_j} is the value of the j-th feature in the i-th syllable. In the same way, weighted standard deviation, when considered, is computed as

$$\bar{\sigma}_j = \sqrt{\sum_{i=1}^{N} w_i (f_{i_j} - \bar{\mu}_j)^2} \tag{3.3}$$

Aside from weighted mean and standard deviation, this step also introduces in the features set speech rate, computed as the ratio between the number of syllables and the speaking time. Speech rate has been shown to be an important emotional marker, together with pitch variability, in Breitenstein et al. (2001). We also include the percentage of time in which vowels are realized $\%V$, computed as the percentage of speaking time occupied by the nuclei associated with each *phonetic syllable*, and ΔC, computed as the mean length of the segments obtained by merging the length of each syllable's head with the length of the preceding syllable's coda. $\%V$ and ΔC were introduced in Ramus et al. (1999) in an attempt to classify languages by their rhythm and, while their effectiveness in this task has been repeatedly questioned, especially in recent years (Arvaniti, 2009), they are still an effective correlate of speech rhythm in general terms. Lastly, the maximum among the energy peaks associated with syllable nuclei is retained. A summary of the final *prosodic* features set is presented in Table 3.1 while a block diagram of the features extraction process is represented in Figure 3.1.

	Max	Mean	Standard deviation
Nucleus length		✓	✓
Syllable length		✓	✓
F0		✓	✓
Speech Rate		✓	
Harmonicity		✓	✓
Energy	✓	✓	✓
Shimmer		✓	
Glissando likelihood		✓	
TEO		✓	✓
ZCR		✓	
ΔC		✓	
%V		✓	

Table 3.1: Statistics computed over the features extracted from the syllables in the utterance (prosodic only). %V and ΔC are considered as mean values over the whole utterance.

3.3 Emotion regression

In this section we report an analysis of the features' predictive power and the performance of the automatic emotion regression task for a qualitative and quantitative view on the obtained results. During the features extraction step, one file containing a very short utterance was removed from the corpus as the syllabification algorithm was not able to detect any *phonetic syllable* in it.

3.3.1 Material

In this experiment, the Vera am Mittag (VAM) corpus (Grimm et al., 2008) is used. The material consists of 48 minutes of audio recordings from a German talk show. The recordings were manually annotated by a pool of human judges with continuous values

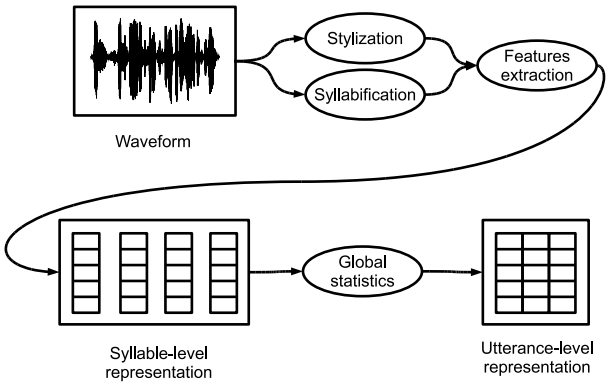


Figure 3.1: Blocks diagram of the features extraction process.

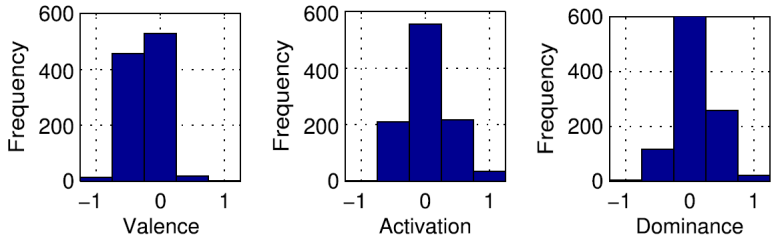


Figure 3.2: Scores distribution for the three axes in the VAM corpus as reported in Grimm et al. (2008)

in the VAD space. The VAM corpus is divided into two subsets. The first (VAM I) is composed by 478 utterances from 19 speakers assessed by 17 human evaluators while the second (VAM II) is composed by 469 utterances from 28 speakers assessed by 6 evaluators. For our tests, the full corpus (VAM I+II) is used. Due to the material contained in the corpus, while Activation and Dominance scores are quite balanced, the Valence scores are unbalanced towards negative values, as shown in Figure 3.2.

3.3.2 Features analysis

To evaluate the raw predictive power of the *prosodic* features with respect to each of the considered axes, we computed the Spearman's ρ and found that all the features are correlated with at least one axis in a statistically significant way. Detailed graphs reporting

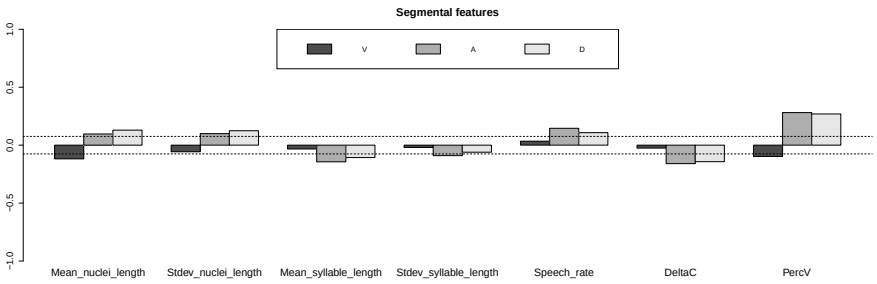


Figure 3.3: Spearman's rho *for segmental features with respect to the three axes. The dotted line shows the critical value for* $\alpha = 0.01$.

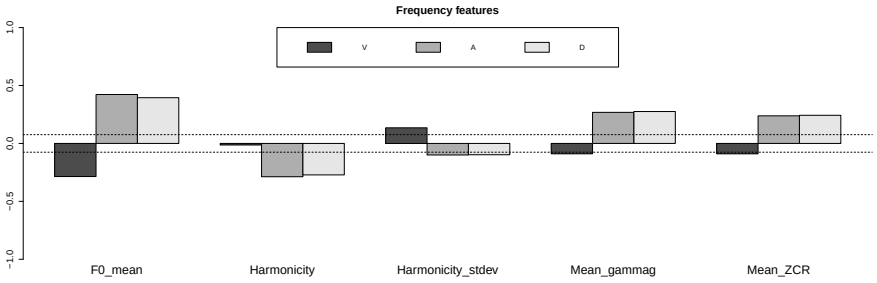


Figure 3.4: Spearman's rho *for frequency features with respect to the three axes. The dotted line shows the critical value for* $\alpha = 0.01$.

the actual ρ values with respect to a threshold of 0.075, representing the critical value for 944 degrees of freedom and a significance level of 99%, are reported in Figures 3.3, 3.4 and 3.5. Each graph shows the correlation values for a subgroup of the *prosodic* features, divided in *segmental*, *frequency* and *energy* related.

It is necessary to evaluate the inter-correlation of the features, other than the direct correlation of the features themselves, to assess the predictive power of the obtained representation. To this purpose, the Correlation-based Feature Selection (CFS) algorithm (Hall, 1998) is used. The CFS algorithm selects a subset of the initial features set exhibiting high correlation with the target class or value while keeping low intercorrelation between features: if two features exhibiting high correlation with the target are also strongly inter-correlated, only the feature exhibiting the best correlation with the target value is retained. A Leave-One-Out cross-validation (LOO-CV) test with the CFS algorithm was

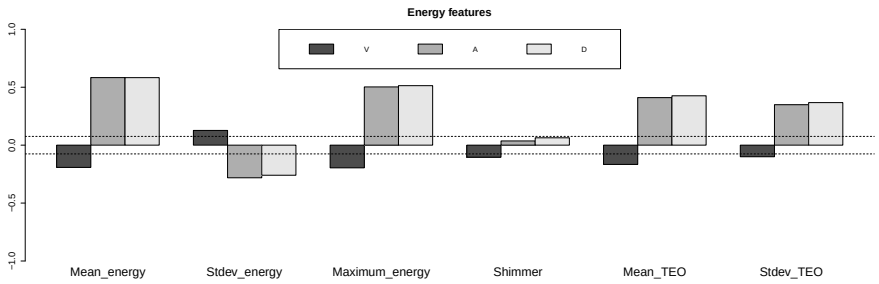


Figure 3.5: Spearman's rho *for energy features with respect to the three axes. The dotted line shows the critical value for* $\alpha = 0.01$.

performed on the features set for each dimension to check which features were selected by the algorithm and evaluate the actual predictive power of the whole features set. Results of this test are reported in Table 3.2. The indications of the CFS algorithm were used only to evaluate the features set: no features were removed when performing emotion regression to avoid overfitting on the VAM corpus.

3.3.3 Results

Automatic prediction, for a voice stimulus, of the coordinates it will have in the VAD space is performed by means of a Support Vector Regressor (SVR). For the presented tests, the LibSVM implementation (Chang and Lin, 2011) of an ϵ-SVR with a Radial Basis Function (RBF) kernel was used. This is the same kernel used in Wu et al. (2010). As in the reference work, absolute error is used as performance measure together with Pearson's correlation coefficient, computed as:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3.4}$$

In Wu et al. (2010), the complexity parameter C and the γ parameter of the RBF kernel were optimized on the whole training set by means of a grid search. This was because the test protocol consisted of a LOO-CV to allow comparison with the results

Table 3.2: CFS results for prosodic features. For each feature and each dimension, the percentage of times the feature was selected by the algorithm in a LOO-CV setup to predict the value of each dimension is reported.

	Valence	Activation	Dominance
Mean nuclei length	100%	0%	0%
Stdev nuclei length	100%	0%	0%
Mean syllable length	100%	0%	0%
Stdev syllable length	8%	5%	0%
F0 mean	100%	100%	0%
Speech rate	100%	0%	0%
Harmonicity	0%	100%	100%
Stdev Harmonicity	100%	0%	0%
Mean energy	100%	100%	100%
Stdev energy	100%	0%	0%
Maximum energy	0%	100%	100%
Shimmer	100%	0%	0%
Mean Γ_g	100%	100%	100%
Mean TEO	100%	100%	100%
Stdev TEO	100%	100%	100%
ΔC	1%	100%	100%
%V	100%	100%	100%
Mean ZCR	100%	0%	100%

Table 3.3: Pearson correlation coefficients and absolute errors obtained by the SVR both in the 10-Fold-CV setup and in the LOO-CV setup. Results obtained in Wu et al. (2011) are reported for comparison.

	Pearson correlation coefficient					
	10-Fold-CV			LOO-CV		
	Valence	Activation	Dominance	Valence	Activation	Dominance
PROS Only - Wu et al.	-	-	-	0.47	0.77	0.75
PROS + MFCC - Wu et al.	-	-	-	0.56	0.83	0.80
PROS Only - Proposed	0.37	0.77	0.75	0.39	0.78	0.76
PROS + MFCC - Proposed	0.45	0.80	0.78	0.48	0.81	0.79
	Absolute error					
	10-Fold-CV			LOO-CV		
	Valence	Activation	Dominance	Valence	Activation	Dominance
PROS Only - Wu et al.	-	-	-	0.13	0.17	0.15
PROS + MFCC - Wu et al.	-	-	-	0.12	0.15	0.14
PROS Only - Proposed	0.14	0.17	0.16	0.14	0.17	0.15
PROS + MFCC - Proposed	0.13	0.16	0.15	0.13	0.16	0.14

reported in Grimm et al. (2008) and the authors assumed that a single sample would not have had a significant impact on the choice of parameters. The same was assumed in the Feature Selection (FS) step they included in their approach. In this work, we present results obtained both with 10-Fold cross-validation (10-Fold-CV) and LOO-CV. While in the LOO-CV tests we used the optimal parameters found for the entire training set by means of a grid search, the results obtained with the 10-Fold-CV protocol were obtained by optimizing the parameters on each of the ten training sets by internal 10-Fold-CV and then evaluating the obtained model on the test fold. Final results, together with the reference ones, are reported in Table 3.3.

3.3.4 Discussion

In this experiment, I concentrated on a dimensional representation of emotions by performing regression in the VAD space as, in this case, I prefer a top-down approach starting from a continuous emotional representation to reduce the influence of emotional words and work later on the identification of areas in this space to which emotional labels can be assigned. While this is mainly useful for communication needs it can also be used, as I will show in Chapter 4, to define part of the behavior of a robotic architecture.

Concerning the choice of the unit of analysis, this work focuses on the use of the *phonetic syllable*. Syllables were used in previous works in the automatic emotion classification task (i.e. Kao and Lee (2006)) but they are typically obtained by Forced Alignment (FA) or Automatic Speech Recognition (ASR) modules, consequently assuming the presence of a trascription or, at least, of a dictionary, which is limiting from the point of view of technological applications. ASR, in particular, is only justified, in my opinion, if linguistic features are to be considered for emotion recognition, which is a different area of interest than the one I am exploring here.

In this work, I avoid the need of FA or ASR by using a phonetic definition of syllable so that all we need to take into account is the structure of the signal in terms of energy profile and voicing information. The use of a phonetic, rather than a phonological, definition of the analysis unit introduces, of course, a number of problems related to syllabification like insertion and deletion errors, where the former is caused by syllable splitting by the occurrence of artifact energy peaks and the latter is caused by syllables merging typically in cases of coarticulation. As I have shown in Chapter 2, these problems usually lower performance measures of syllabification algorithms when their output is compared with a manual transcription taken as reference and usually following phonological segmentation rules. For features extraction, however, it is not a problem to work on a segmentation into syllable-like units that only partially overlaps with manually marked syllables as a segmentation of the signal into acoustically self-consistent units is sought rather than a correspondence with phonological expectations. As an element of novelty with respect to

other approaches using units of analysis with length ranging from frame-level to word-level, including syllables, I am not considering all areas of the signal in an equal way. By extracting energy and periodicity related features together with MFCCs from the *phonetic syllables* nuclei only, we discarded all spectral information related to consonants, which mostly introduces noise. This approach is motivated by studies on the acoustic properties of emotional speech that concentrated on vowels only. In Drioli et al. (2003), the voice quality characteristics of stressed and unstressed vowels pronounced in VCV setting were investigated, in Patel et al. (2011) the correlation with a set of discrete emotions of a number of acoustic features extracted from sustained \\a\\ sounds was explored while in Vlasenko et al. (2011) and in Gharavian et al. (2012) the particular importance of formants when distinguishing emotions has been reported.

Another point of novelty in this work is represented by features weighting in terms of nuclei durations. By weighting syllable-level features on the basis of the relative portion of nuclei time they occupy, more importance is assigned to long nuclei while reducing the influence of shorter nuclei. Since duration is the main cue when detecting prominent syllables, this approach is in line with the indications given in Seppi et al. (2010) that features extracted from prominent syllables perform in a similar way to features extracted at word level and contain better emotional markers than their non-prominent counterparts.

An obvious advantage of limiting the extraction of spectral features from *phonetic syllables* nuclei only is the amount of data that needs to be analized in order to obtain performances comparable to the state of the art. Since in Wu et al. (2010) a Voice Activity Detection module was used to avoid extracting features from silent areas, we can estimate that the amount of speech processed in the reference work was roughly equal to the time covered by our automatically detected syllables, which is 36 minutes. Since automatically detected syllable nuclei cover a total of 22 minutes, these results are obtained by processing 40% less of the signal.

From the data shown in Table 3.3, we can observe that performance equals state of the art on Activation and Dominance while it is lower on Valence. Low performance of automatic emotion labeling systems on Valence is frequently found in the literature

when acoustic features only are involved: the opposition between anger and happiness is typically reported to be the most difficult to model. While obtained results are higher than the ones reported in Grimm et al. (2008) (where a mean correlation equal to 0.6 and a mean absolute error equal to 0.24 were reported), the approach presented in (Wu et al., 2010) is better than the one I am presenting on this particular dimension, although even in that case performance is still low. A first observation we can make about this is that the main difference between our approach and the reference one lies in the lack, in our features set, of descriptors of spectral dynamics as in (Wu et al., 2010) deltas and double deltas of MFCC coefficients were included. Not having this kind of information appears to have no effect for what it concerns Activation and Dominance regression while influencing Valence regression only. In the literature, it has been repeatedly found that static classifiers obtain acceptable performance on the reference corpora and this has led to the wide popularity of SVMs on emotion classification. However, our results, combined with the results presented in reference work, seem to suggest that, while this approach works in an acceptable way on Activation and Dominance, it may be necessary to adopt a different strategy for the Valence axis only. This possibility may have been masked in the past by posing the emotion recognition problem in terms of classification. Having a single classifier working on a full set of classes, in fact, does not allow to separate the process dedicated to separating emotion classes that are distinguishable only in terms of Valence. On the contrary, in the emotion regression problem, by having a regressor dedicated to each axis, it is easier to structure a system tracking emotional levels on Activation and Dominance using static approaches based on global statistics while having a separate regression algorithm tracking Valence by taking into account features dynamics.

From the qualitative analysis of the features performed on the VAM corpus, both in terms of absolute correlation with the VAD axes shown in Figures 3.3, 3.4 and 3.5 and in terms of choice of the CFS algorithm, we obtain useful insight. First of all, we observe that ΔC appears to be dominating speech rate. The two are strongly correlated (ρ: -0.56), in line with the basis on which the $Varco\Delta C$ rhythmic measure (Dellwo, 2006) has been developed and with the findings reported in Barry et al. (2003); Dellwo and Wagner (2003).

ΔC, however, has better correlation with Activation (ρ= -0.16 vs ρ= 0.15) and Dominance (ρ= -0.14 vs ρ= 0.11). This is confirmed by the CFS data for Activation and Dominance while for Valence the CFS algorithm preferred speech rate although neither of the two was significantly correlated with that axis. CFS appears to be selecting speech rate only because there are almost no features correlated with this axis in a strong way.

In (Wu et al., 2010), it was assumed that the LOO-CV setting has no influence on the choice of the features: while this is not entirely true, as in Table 3.2 some features were occasionally selected by the algorithm, in most cases if a feature was selected once, then it was included in every features set. This result validates Wu's assumption and supports our choice to use the optimal parameters found on the full dataset for the SVM in LOO-CV experimental setup. This choice is also supported by the comparison between the results we obtained in the 10-Fold-CV setup with the ones obtained in LOO-CV setup, in which only a minor performance difference in terms of correlation coefficient can be observed on the Activation axis.

CFS features selection consistently discards segmental features, with the exception of ΔC and $\%V$, thus suggesting that their role is exhausted in features weighting. This is in line with what it has been shown in (Seppi et al., 2010) regarding the particular importance of the acoustic content found in prominent syllables and it validates our attempt to introduce an adaptive way of considering features along the speech signal, avoiding to consider them as equally important in every area. Since our method of features weighting is based on nucleus duration, which represents the primary cue for prominence detection, this observation also opens the way to the introduction of features weigthing based on prominence scoring for emotion regression. This, in our opinion, would represent an important step to recognize important areas of the speech signal, allowing an automatic system to focus on them to obtain cleaner data.

Lastly, the Γ_g feature appears to be a good descriptor for pitch, being always selected by the CFS algorithm. This is interesting as the Γ_g value represents a continuous score for the occurrence of glissandos in the utterance and the correlation of this particular phenomenon with emotions, to our knowledge, has never been tested before.

3.4 Continuous emotion regression

I will now concentrate of continuous emotion regression on a per-syllable basis. For the following tests and for the affective robotics system I will present in Chapter 4. As the current interest is to evaluate the contribution of a single unit to emotion tracking, I will limit my analysis to the correlation of a single syllable's features vector with the instantaneously perceived emotional level. I will then show how this analysis allows to set up an affective robotic architecture working in real-time with syllable based analysis. While it is possible to enlarge the features' scope by employing features vectors from the preceding syllables, this is left for future work as it makes more complicate to present a clear interpretation of the contribution of each single feature for continuous emotional tracking.

3.4.1 Material

For the test presented in this Section, the SEMAINE corpus (McKeown et al., 2010) is used. The considered dataset is composed of 55 interaction sessions between human subjects and artificial sensitive listeners each one having a distinct personality (happy, gloomy, pragmatic...). From the audio recordings of the human subjects involved in the interactions, silent intervals long at least 2 seconds were marked as separators between isolated utterances. With this method, 882 segments were extracted for a total speaking time of 3 hours and 28 minutes.

In the SEMAINE corpus, continuous annotations are available on the three axes we are considering in this Chapter, among others. By considering the mean instantaneous scores provided by the human judges, a single stream of continuous values is obtained. The number of raters per session varies from a minimum of two to a maximum of 8, with the majority of the sessions having 6 raters. After applying the segmentation strategy presented in Chapter 2 and filtering out syllables having nuclei less than 64ms long, 29440 *phonetic syllables*, 75% of the total number of syllables found, were extracted. As, in

Table 3.4: Pearson correlation coefficients and absolute errors obtained by the SVR on the SEMAINE corpus.

	Pearson correlation coefficient		
	Valence	Activation	Dominance
PROS Only	0.27	0.47	0.22
PROS + MFCC	0.4	0.57	0.31
	Absolute error		
	Valence	Activation	Dominance
PROS Only	0.22	0.2	0.16
PROS + MFCC	0.2	0.19	0.16

this experiment, we are interested in evaluating the predictive power of each automatically detected syllable, the target value associated with each unit is computed as the mean value of the scores inside that unit.

3.4.2 Results

As in Section 3.3, an SVR was trained for each axis. This time, this is done on the basis of each automatically detected syllable. The target variable is computed by taking first the mean value assigned by each human judge to each frame and then considering the mean value over the entire syllable for each axis. Table 3.4 shows a summary of the obtained results. Since using *phonetic syllables* as units for continuous emotion recognition on the SEMAINE corpus represents a novel approach, there is no directly comparable work to take into account. As in Section 3.3, I report the results obtained both with prosodic features only and with MFCCs. Correlation coefficients for the considered prosodic features are shown in Figure 3.6.

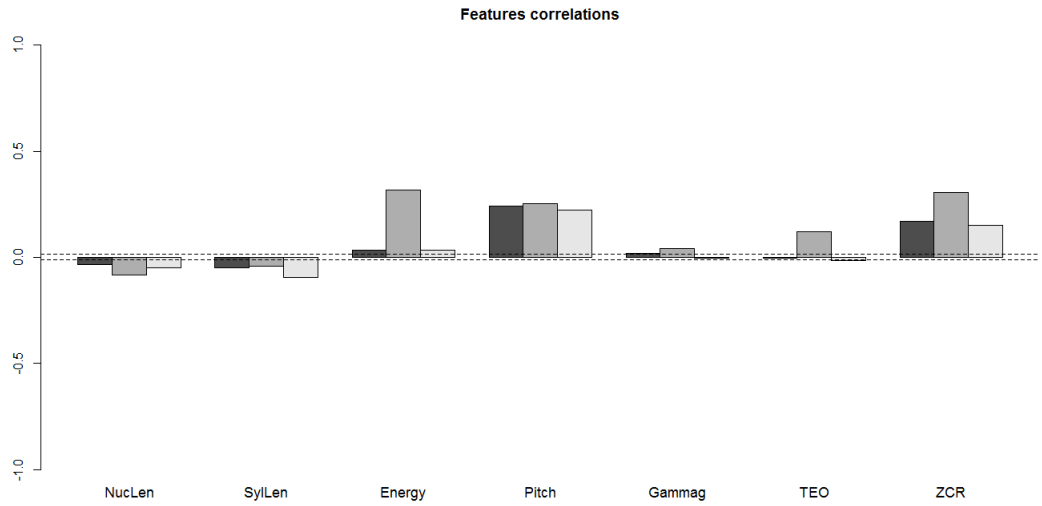


Figure 3.6: Spearman's rho of the considered features with respect to the three axes on a per-syllable basis. The dotted line shows the critical value for $\alpha = 0.01$.

3.5 Conclusions

I presented a system for automatic emotion regression in the VAD space based on *phonetic syllables*. I have shown that state of the art results for Activation and Dominance can be obtained by extracting spectral features from *phonetic syllables* nuclei only and weighting them by the relative duration of each nucleus. As a performance drop with respect to the reference approach can be observed only in the case of the Valence axis, on which state of the art performance is still low, it can be hypothesized that, while a static regressor may be successful for Activation and Dominance, dynamics may be important for the Valence axis only. Should this be the case, a dimensional model of emotions would have to be preferred to a discrete one, from a technological point of view, because it allows to isolate and easily deal with a problematic area of emotion recognition, often represented by the difficult problem of modelling anger/happiness opposition.

I also presented a qualitative analysis of the considered features showing that segmental features appear to exhaust their role in features weighting, with the exception of rhythm-related features, opening the way to the introduction of prominence scores into features

extraction and, together with the fact that state of the art results can be matched at least on Activation and Dominance regression by considering only syllable nuclei for spectral information, highlights the need to avoid considering all areas of the speech signal as having the same importance for emotion regression. From the same qualitative analysis, we have shown that three not commonly used features, ΔC, $\%V$ and Γ_g, appear to be powerful descriptors of the emotional content.

The final test on continuous emotion tracking, performed on a per-syllable basis, shows that instantaneous emotional tracking with no contextual information appears to be reliable for the Activation axis only. From the reported analysis, relevant prosodic features for this task appear to be the mean pitch and energy, the nucleus length, the TEO and the ZCR.

Chapter 4

Emotional speech driven robotic architecture

Emotions are tightly bound to the physical world and should not be treated as abstract classes. Moreover, it is widely accepted that obtaining ground truth labelings of audio and video recordings is a difficult task. This is because emotion perception can vary a lot among judges, making the final scores less reliable. This difficulty comes from the fact that, while we are perfectly aware of what we are talking about when we discuss emotions, it is not as simple to define emotions. Following the theory presented in Ledoux (1998), as I discussed in Chapter 1, I follow the idea that the word *emotions* does not refer to something that actually exists in the brain. Emotions are linguistic tricks that allow us to describe complex physiological experiences. For this reason, while using emotional corpora is indeed useful to study emotions, it is necessary to evaluate the capability of an automatic system of recognizing emotions by its capability of reacting accordingly to the emotional stimuli coming from a user. This way, it is possible to verify the performance of the emotional speech analysis module by considering the consistency of the robot's behavior. In this Chapter I present a general affective robotics architecture in which I include a real-time implementation of the syllable based speech analysis method

presented in Chapter 2 and tested on emotional speech corpora in Chapter 3. I will consider the Activation axis only as it appears to be the most reliable given a context of one *phonetic syllable*. I will also not include the trained SVM model to perform emotion tracking. This is because the language the models were trained on was English and I only had Italian speakers available and because my goal is to show that, by keeping an interpretable features set, it is possible to better illustrate how the robot works. Specifically, even though the presented task is simple, I believe that, if a certain system works, it is important to be able to explain *why* it works. A linguistically motivated approach to speech analysis is helpful, in this sense, as it allows the system to be described by including terms coming from a well-established terminology concerning prosody.

4.1 Virtual creatures

Robotic architectures simulating real creatures, both animal or humanoids, are often referred to as *virtual creatures*. From a commercial point of view, many of these can be considered the rightful successors of the old tamagotchi concept: they are designed to simulate the basic need of pets like feeding and caring and to acts like artificial companions.

Among recent approaches to this kind of product, the Aibo robot, shown in Figure 4.1, represents a very well known example. Aibo was developed by Sony and distributed between 1999 and 2005. The latest model was characterized by a MIPS R7000 processor, 64MB of RAM and 20 degrees of freedom distributed among legs, head and tail. It was controlled by a proprietary operating system developed by Sony, named Aperios, featuring a modular architecture to allow easy integration with modules sold separately. While designed with the primary objective of being a commercial entertainment product, Aibo has been also used in academic research to study, for example, children's behaviour (Batliner et al., 2011) and verbal human-robot interaction (Kuremoto et al., 2011).

The Nao robot, shown in Figurefig:Nao, is a widely used humanoid platform designed mainly for research purposes that allows users to experiment both with locomotion and with human-robot interaction. It features 25 degrees of freedom, 2 cameras, 4 microphones,

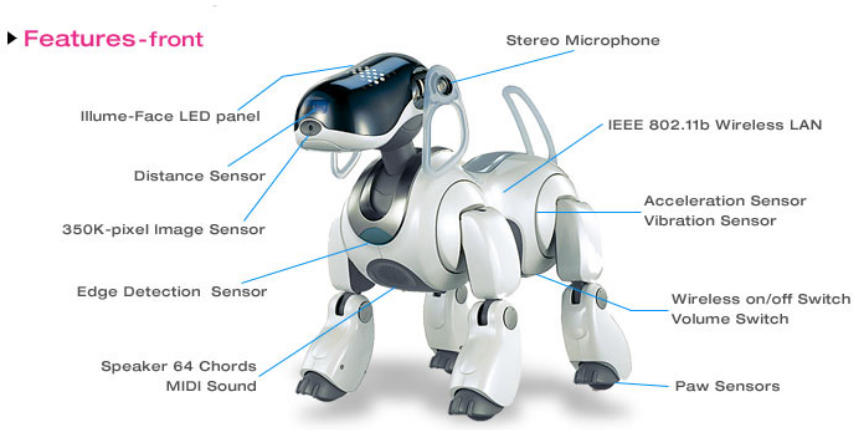


Figure 4.1: The Aibo robot

9 tactile sensors and 8 pressure sensors. It is equipped with an Intel ATOM processor and runs a Linux kernel operating with a proprietary middleware (NAOqi). While it does not come with a built-in personality, Nao is often provided by researchers with basic communication capabilities and it is used as a virtual companion in human robot interaction therapy for autism (Shamsuddin et al., 2012) and to study emotional movements in social games (Barakova and Lourens, 2010).

In this work, I use the Pleo robot, shown in Figure 4.3. While being a relatively cheap commercial product, Pleo is provided with a simple programming interface based on the PAWN language that allows users to write original behaviors for the robot overriding its pre-defined artificial personality. It has 14 degrees of freedom, an Atmel ARM7 processor, 12 touch sensors, two binaural microphones and a camera dedicated to light detection and basic navigation and object tracking. Given the limitations of the PAWN programming language and of the available hardware, I took advantage of the USB interface to perform signal processing on a dedicated PC and delegated the necessary animation control to a pre-loaded PAWN script. The Pleo robot was chosen for the presented experiments because, by representing a pet, I assumed it would have been more easily accepted by the human subjects although the simulated intelligent behaviors are limited.

Figure 4.2: The Nao robot

4.2 Proposed architecture

The affective robotics architecture presented here is used to discuss the following observations:

Observation 9. *using emotions can aid the design of a modular robotics architecture by acting as an interface between perception and action, thus abstracting the low-level decision processes from the raw signal processing phase*

Observation 10. *a linguistically motivated method to process speech, such as the one presented in Chapter 2 has an impact on the performance of a technological system*

Observation 11. *it is possible to test a continuous emotion tracking system by* using *it in a simple task rather than by comparing results with manual annotations, that are only partially reliable*

Concerning Observation 9, as it happens for humans from a linguistic point of view, emotions can be used in robotics architectural design to abstract complex configurations in

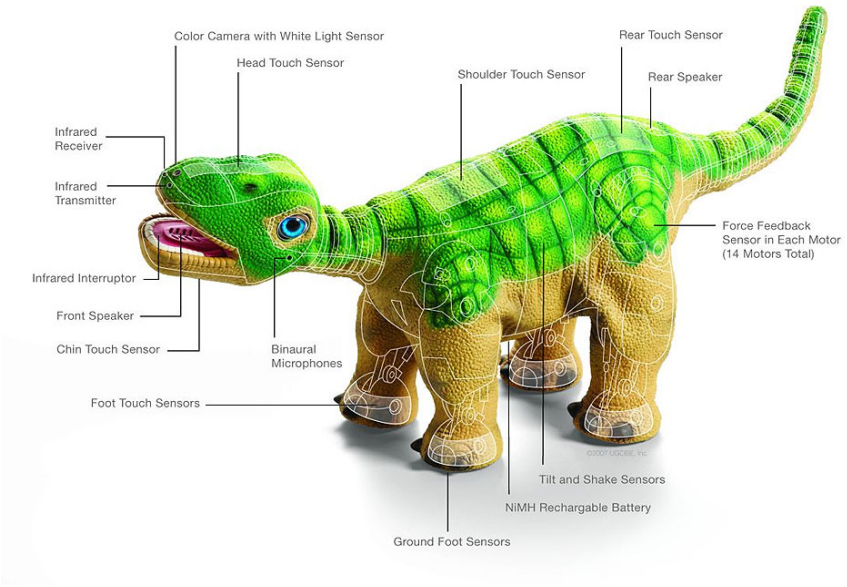


Figure 4.3: The Pleo robot

a simple way. This abstraction can be then used in order to let a technological approach react accordingly to the experienced synthetic emotion. By representing internally the result of the evaluation of all the channels over which the synthetic emotion is computed, decisions can be taken on the basis of a summary of the results rather than considering every single channel independently. In the architecture used to test the syllable-based emotional speech tracker, represented in Figure 4.4, I define an interface based on instantaneous emotional stimuli coming from signal analysis modules. The internal model represents the current emotion in terms of the four dimensions described in Fontaine et al. (2007) (Activation, Valence, Dominance, Unpredictability) so incoming emotional stimuli are represented by 4-dimensional vectors. Although in the implementation used in this work only the module dedicated to speech analysis is connected to the emotional interface, the emotional model is designed to compute the emotional impulse on each of the four axes by taking the mean value of all the incoming pulses. As not every signal processing module may be designed to influence the emotional state on all the considered axes, modules that do not intend to contibute in defining the emotional state on a specific axis are allowed to transmit a *Not-a-Number* (NaN) component in the 4-dimensional impulse vector. This way, the emotional

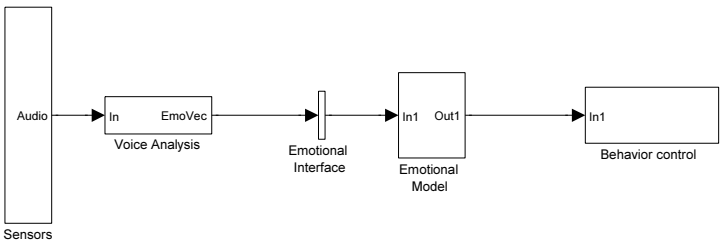


Figure 4.4: Generic model of the affective robot

model does not consider the contribution of the module when computing the mean emotional impulse on that axis. In the presented experiments, the real-time speech processing module sends emotional pulses on the Activation axis only. This is because, by looking at the data presented in Chapter 3, this is the axis about which we are able to obtain the most reliable results. The Voice Analysis module shown in Figure 4.4 is designed to send pulses in the form o a 4-dimensional vector containing $[x, NaN, NaN, NaN]$ where x represents an emotional impulse on the Activation axis. This design choice is intended to leave the architecture open to allow an easier expansion in the future by including more channels. Figure 4.4 also shows a decision module working on the basis of the internal emotional state. This highlights how the decision process is not concerned with signal processing but relies only on the internal configuration resulting from the perception step and abstracted in terms of synthetic emotion. Also, the emotional model works in continuous (e.g. dimensional) mode only. Discretizing the dimensional representation, if necessary, is left to decision modules depending on what they are designed to do. Details on the decision module used in this implementation of the emotional architecture are described in Section 4.3.

Concerning Observation 10, it is necessary to describe how the offline speech processing method presented in Chapter 2 has been moved to a real-time environment. The most important implication of the speech processing method described in this work is that a linguistic unit, the *phonetic syllable*, is considered as the basis on which emotion tracking is performed. This is the case for the real-time system too. Instead of providing emotional

pulses on a constant frequency (i.e. with a frame-based strategy), these are sent to the emotional model on the basis of the variable frequency depending on syllable occurrence.

Given the definition of *phonetic syllable* reported in Section 3.1, the phonetic concept of syllable can be introduced in an technological real-time system by using buffering. As the definition describes the syllable in terms of voiced energy peaks, it is necessary to compute on a frame basis only pitch and intensity of an incoming signal. In the presented system, instantaneous pitch and intensity measures are accumulated in a buffer until the template of a phonetic syllable is detected. When this condition is verified, the syllable is passed to the features extraction module and an emotional impulse is generated on the basis of the method described in Chapter 3. This means that, in the presented system, the features extraction module is not always active. On the contrary, it is disabled, thus reducing the computational load, until a *phonetic syllable* is detected. Moreover, as spectral features are computed over the syllable nucleus only, the system saves the effort needed to examine non-nuclear portions of the signal. As shown in Chapter 3, this does not alter significantly the performance on the emotion regression task.

Figure 4.5 shows the frame level computation module dedicated to buffering. Given the incoming signal and its Fast Fourier Transform (FFT), the intensity value of the frame is computed and, if the silence threshold is exceeded, autocorrelation based Voice Activity Detection is used to verify if the frame is voiced and, if that is the case, what is its pitch. Intensity, voicing, and frequency information, along with the raw signal, are saved in dedicated buffers.

Figure 4.6 shows how the frame based analysis module allows the use of syllable based analysis in real-time. From the intensity buffer, the presence of a local maximum not yet associated with a syllable is verified. If a voiced local maximum followed by a local minimum is found in the buffer and it has not been already assigned to a syllable, the template check module enables the features extraction module. By considering the buffered data, the syllable-based module generates an emotional impulse in the form of a 4-dimensional vector, sending it to the emotional model to be processed.

Figure 4.7 shows how the system extracts the last syllable found in the buffer. At this

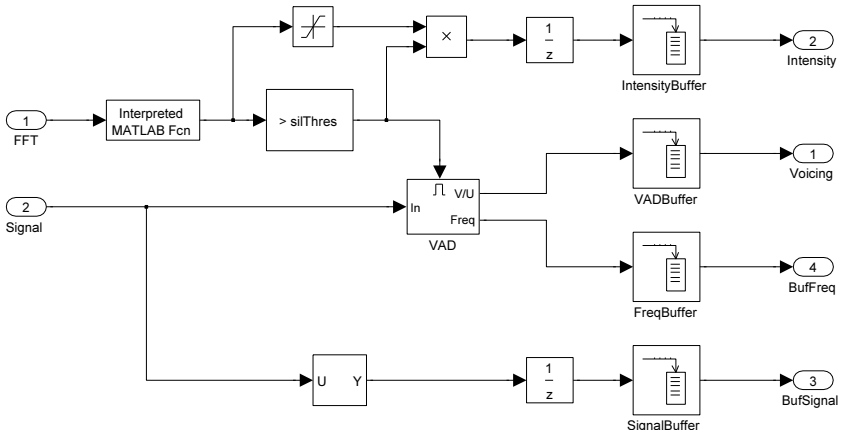


Figure 4.5: Intensity computation, Voice Activity Detection and buffering of the considered channels.

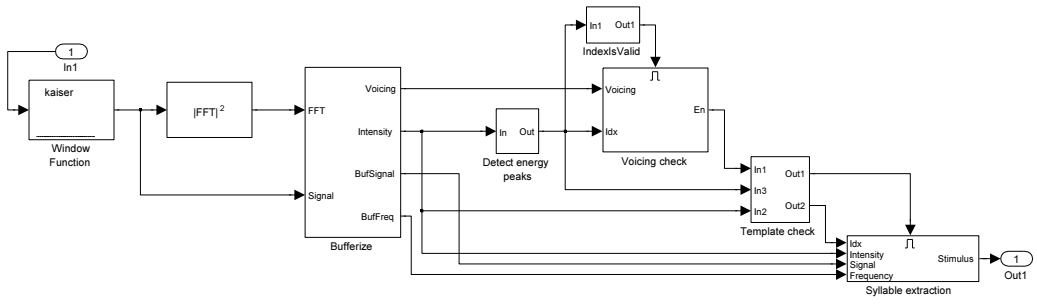


Figure 4.6: Buffering and syllable template detection

Table 4.1: Features used in the real-time system. The correlation coefficient computed on the SEMAINE data and the adjusted weight used to compute the emotional stimulus are also reported.

Feature	Correlation	Weight
Nucleus length	-0.08	-0.1
Nucleus energy	0.29	0.3
Mean pitch	0.25	0.2
TEO	0.08	0.1
ZCR	0.3	0.3

stage, the length of the nucleus is extracted and the emotional stimulus is computed only if the nucleus is long enough to contain clear spectral information (during the experiments this limit was set to 80ms). The emotional stimulus is computed as the weighted mean of a subset of the prosodic features presented in Chapter 3 by taking into account the correlation coefficients of the considered features with the manual annotations of the SEMAINE test (per-syllable emotion tracking) Since the the real-time pitch tracking module was less reliable than the offline pitch tracker used in the experiments with emotional corpora, the value of the *mean pitch* coefficient was slightly reduced in favor of the other features. The considered features along with their original correlation coefficient and the adjusted weight is reported in Table 4.1.

The emotional model, shown in detail in Figure 4.8, keeps track of the internal emotional state of the robot. It receives emotional stimuli from the signal processing modules and computes the new emotional state on the basis of the previous one and by taking into account a simulated tendency to return to a neutral state in absence of emotional stimuli. The emotional stimulus is computed as the mean of the incoming emotional stimuli on each axis. Since in the presented implementation only the voice analysis module generating pulses on the Activation axis is present, the emotional stimuli are computed as the mean between the output of this module and of the back force simulating the tendency towards a neutral state. Back force is computed by multiplying the preceding emotional value of each axis by a constant factor θ ($\theta = -0.0005$ in the presented experiments). After taking

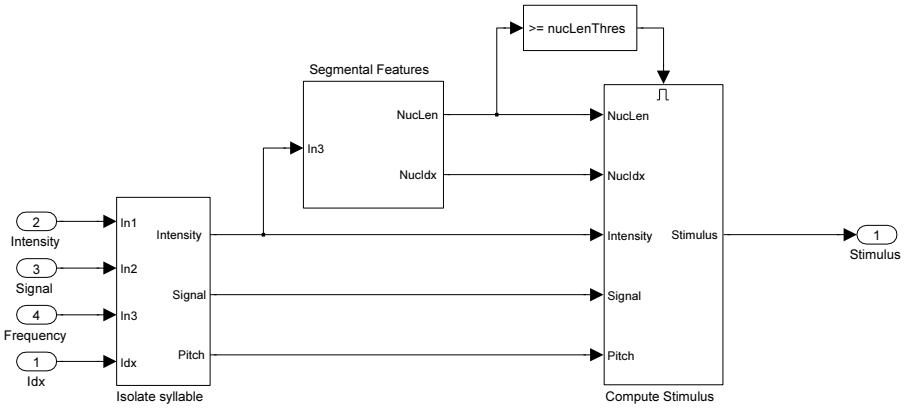


Figure 4.7: Syllable extraction module

the mean value of the contributing factors on each axis, the impact of the stimulus on the synthetic emotion is modulated depending on whether the stimulus is concordant with the present emotional state or not. If the robot is in a positive (negative) emotional state and a positive (negative) stimulus is incoming, the positive (negative) effect is dampened proportionally to the strenght of the present emotional state, otherwise the effect of the stimulus is tripled. This way, weak stimuli, both positive and negative, have less effect on the robot if it is already in a very *excited* or in a very *relaxed* state, repeated stimuli of the same strength have a decreasingly lower impact as their influence sums up and opposing stimuli rapidly change the emotional state of the robot. These simple rules constitute a model of affective adaptation (Frederick, 1999) modulating the impact of the incoming stimulus on the synthetic emotion. Affective adaptation is the process of weakening of the affective response of a constant or repeated affective stimulus by psychological processes. This means both that the positive effect of a stimulus weakens after some time time (simulated by the back force rule) and that continuous exposure to the same stimulus lowers the importance of the stimulus itself over time (simulated by the stimulus modulation rule).

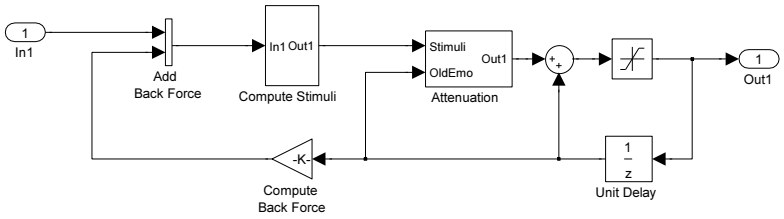


Figure 4.8: The emotional model

4.3 Case study

In order to illustrate how the presented architecture can be deployed and to evaluate the performance of the syllable based emotional speech tracker, a simple task with the Pleo robot has been designed. Subjects participating to the experiment were asked to draw Pleo's attention on a screen when a butterfly was visible and to calm him down when the butterfly was not there. They were told that they were allowed to use voice only to control Pleo and that, although it was listening to them, the robot would not have been able to understand what they were saying so they should treat it like a real cub. To perform signal processing and to remotely control the Pleo robot, a notebook running Windows 7 (32 bit OS, 4GB RAM, Centrino 2 processor) was used in the tests. The PC is connected with Pleo by the USB interface and to a second monitor showing images of a lawn with a flying butterfly or without the butterfly to elicit reactions from the user. The user's voice was captured by means of a Sennheiser headset connected to the PC (sampling frequency was set to 16000). The test runs for 1 minute and the experimental setup is shown in Figure 4.9.

The behavior control module, shown in Figure 4.10, shows the structure of the decision process based on the internal emotional state only. First of all, the first component of the emotional vector, representing Activation, is extracted. The continuous Activation space, normalized in the interval [-1, 1], is discretized into three intervals $A = [-1, -0.5], B =$

Figure 4.9: Experimental setup

$[-0.5, 0.5], C = [0.5, 1]$. Decisions are taken depending on the transition from one area to the other. To avoid multiple transitions if the Activation curve moves around the 0.5 and -0.5 boundaries, the areas are *blended* in the intervals [-0.6, -0.4] and [0.4, 0.6]. A transition from area A to area B is therefore accepted only if the curve crosses -0.4 while rising and a transition from area B to area A is accepted only if the curve crosses -0.6 while descending. The same holds for areas B and C. If an area transition is detected, a command is generated by taking into account the areas involved in the transition. While the generic model running on the control PC sends command strings remotely through the USB interface, animation and sound playing are handled asynchronously by a PAWN script running on Pleo. Table 4.2 shows a summary of the commands sent to Pleo relatively to each possible situation.

When the experiment starts, the butterfly is present on the screen so the user is expected to get Pleo excited. When area C is entered, the system makes the butterfly disappear from the screen with 5 seconds delay in order to have the user calm down Pleo. When area A is entered, the system makes the butterfly appear again with 5 seconds delay in order to have the user get Pleo excited again. Figure 4.11 shows two interaction examples:

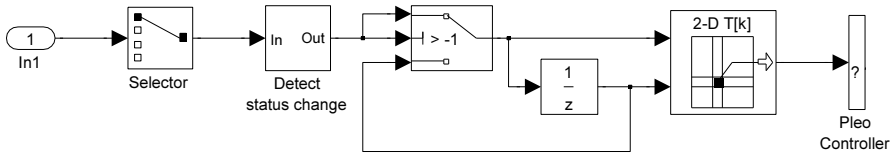
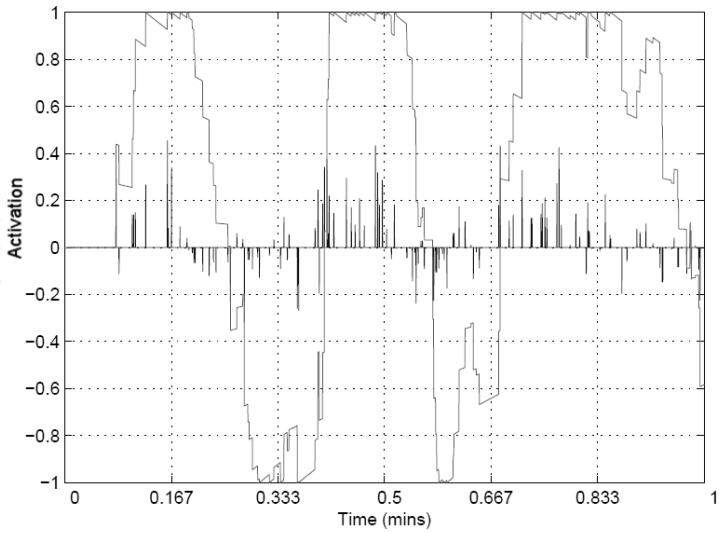


Figure 4.10: Behavior control module

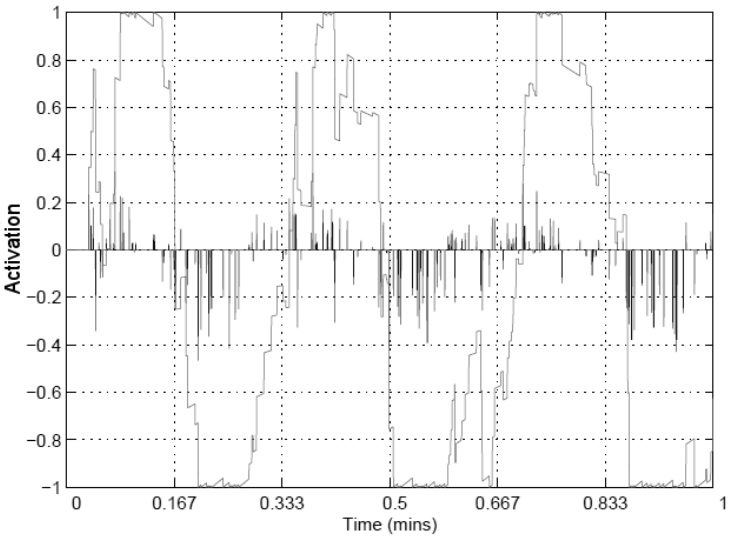
Table 4.2: Command generation matrix.

to area →	A	B	C
A	Keep *relaxed* position	Keep *relaxed* position	-
B	Play *relaxed* animation	Keep last position	Play *happy* animation
C	-	Keep looking at the screen	Keep looking at the screen

black peaks show syllable based emotional impulses while the light grey curve shows how the internal Activation state of the Pleo robot changes as a function of time and emotional stimuli coming from the user's voice. It is possible to observe that, while not being provided with specific instructions about *how* to control the robot, a user is generally able to move the Activation curve in the [-1, 1] interval to produce the expected pattern. 5 female and 4 male subjects were recruited for the experiment. Table 4.3 shows the number of hits for each behavior obtained by each of the subjects. An exhibited behavior is considered a hit only if it was expected. For example, an *excited* behavior is considered a hit only if the butterfly was on the screen and if the Activation curve is kept in the correct area until the butterfly disappears from the screen.



(a) Subject #6



(b) Subject #8

Figure 4.11: Two examples of interaction plots. Emotional stimuli (black peaks) are shown together with the Activation curve (light grey) over time

Table 4.3: Number of hits per expected behavior for the 9 subjects who participated in the experiment.

Subject number	*Excited* hits	*Relaxed* hits
1	2	2
2	3	2
3	2	1
4	1	0
5	2	2
6	3	2
7	3	3
8	1	0
9	3	3

4.4 Conclusions

In this Chapter I have shown an example of how the syllable based speech processing method presented in Chapter 2 and evaluated on emotional speech in Chapter 3 can be deployed in a real-time robotic architecture. By using buffering, it is possible to reproduce the syllable based features extraction method illustrated in Chapter 2 and concentrate the analysis on automatically detected syllable nuclei. This reduces computational load, as the features extraction module is not always active, and introduces linguistically meaningful segmental features in a real-time setting. A generic architecture using emotions as an interface between perception and action has been presented. This architecture is generic in the sense that it does not make assumptions on the number of incoming channels contributing to emotion computation or about the number of behavior controls. The emotional model is dimensional and it considers the four axes (Activation, Dominance, Valence and Unpredictability) indicated in Fontaine et al. (2007). Although the model is four-dimensional, in the presented implementation an example considering Activation only is considered as the predictive capability of the prosodic features set appears to be reliable on this axis only when a per-syllable setup is considered as shown in Section 3.4. By designing a simple

game, it was possible to observe if human users were able to control the behavior of the Pleo robot without being instructed about the system's design. Recruited subjects generally showed a good capability of controlling the robot. The final Activation curve was, as expected, a wave-like pattern in most of the cases. Only two of the considered nine subjects were not able to control the robot. In both cases the subjects were not able to induce the *relaxed* state after reaching the *excited* state. Given the limited number of features considered and the limited context (one syllable) for continuous emotional tracking, this can be considered a promising result.

Conclusions and future work

This work consisted in the development of a linguistically motivated speech analysis method concentrating on intonation analysis for emotional speech recognition. In Chapter 2, I have presented a segmentation algorithm into *phonetic syllables*, the unit I have chosen to be at the basis of my system. I have also presented a pitch stylization algorithm to obtain a perceptual account of the tonal movements and performed a series of experiments on syllabic prominence perception to identify the relevant features allowing the analysis method to differentiate among units on the basis of their perceptual importance. The syllable based analysis method presented here leaves room for improvement by introducing a better description for the *phonetic syllable*. Error analysis revelas that some situations appear to introduce systematical errors in the segmentation when the definition of the considered units fails to detect a boundary. This is mainly due to spectral alterations not being captured by the intensity profile which can be summarized by considering situations where the energy distribution among frequencies varies without changing the total amount of energy in the signal. Extending the definition of *phonetic syllable* may also improve the performance of the SOpS algorithm by making its psychoacoustical basis closer to the original definition of the Spectral Constraing Hypothesis, where generic changes in the spectrum of the signal were considered rather than syllabic subparts.

In Chapter 3, I have shown how the analysis method can be applied to the task of emotional speech recognition, in a dimensional setup. I have also shown how the same method can be applied to continuous emotional speech tracking by keeping the *phonetic syllable* as basic unit. Concerning dimensional recognition, results have shown that per-

formance competitive with the state of the art can be reached by analyzing a significantly lower percentage of the signal. The results obtained on the emotion tracking task, moreover, have shown that, although considering a very limited context of one *phonetic syllable*, the correlation between the automatically obtained values on the considered axes and the manual annotations obtained from human judges is strong enough to allow real-time implementation in a robotic system at least for Activation. Future work concerning emotion recognition on emotional speech corpora will consist of improving performance by extending the context to more syllables. Indications concerning context extension may come from the experiments on prominence perception, where the perceptual relevance of a unit with respect to its surrounding ones was estimated to be equal to three syllables.

In Chapter 4, I have shown how the considered analysis method can be ported in a real-time setting, thus keeping the advantage of being able to describe its functioning through linguistic terminology. The proposed robotic architecture considers emotions as an interface between perception and action. Action modules are not concerned with the analysis of raw data coming from sensors but they rely on an abstraction of the results of this kind of analysis performed by dedicated modules. From a technological point of view, this consists of a modular design in which sensors and analysis modules can be added to the architecture without altering the decision processes. Decision processes, on the other hand, can be extended without altering the way data coming from the sensors are analyzed.

An affective robotic architecture can also be used to get around the *ground truth* problem research has encountered in studying emotions. While using emotional speech corpora has been very useful to design analysis methods and compare them, the need of obtaining reliable human judgments for perceived emotional experience is hard. By designing affective robotics architectures and by proposing tasks that can only be completed if successful emotional communication is realized, it may be possible to better estimate how well the proposed methods perform by taking back emotions from an abstract *annotative* level to the *applicative* level they belong to. User experience, in other words, may be a better performance indicator than the correlation with manual annotation. Designing better tests and a psychologically valid questionnaire to evaluate user experience without explicitly ask

users to talk about emotions will be matter of future works.

Appendix A
The Prosomarker tool

Prosodic research in recent years has been supported by a number of automatic analysis tools aimed at simplifying the work that is requested to study intonation. The need to analyze large amounts of data and to inspect phenomena that are often ambiguous and difficult to model makes the prosodic research area an ideal application field for computer based processing. One of the main challenges in this field is to model the complex relations occurring between the segmental level, mainly in terms of syllable nuclei and boundaries, and the supra-segmental level, mainly in terms of tonal movements. I present here a tool for automatic annotation of prosodic data, the Prosomarker, designed to give a visual representation of both segmental and suprasegmental events usinge the syllabification an pitch stylization algorithms presented in Chapter 2. The representation is intended to be as generic as possible to let researchers analyze specific phenomena without being limited by assumptions introduced by the annotation itself.

Architecture

The system architecture is composed of two main processes running independently. The first one is dedicated to data extraction from the segmental level. This process extracts the energy profile to detect syllable nuclei and position syllable boundaries as shown in Chapter 2. The second process deals with suprasegmental analysis of the speech signal by

means of the SOpS algorithm (see Section 2.5).

The INTSINT coding scheme (Hirst et al., 2000) is then used to produce the automatic annotation. INTSINT was chosen among the various coding schemes because it was specifically designed to annotate the target points of a stylized curve, thus producing a phonetic, rather than phonological, account of intonation. Instead of using the target points of the MOMEL curve (Hirst and Espesser, 1993), Prosomarker uses the target points defining the SOpS curve. While the pitch stylization and annotation process is always performed, segmental analysis and annotation are performed only if the user chooses to visualize this kind of events.

In Figure A.1, the Prosomarker architecture is summarized. The design is modular in order to be easily updated by working separately on the syllabification algorithm and on the pitch stylization algorithm. Modular independence also leaves open the possibility, in the future, to parallelize the process, thus saving computational time, and to extend the analysis. For the implementation, we chose to employ the well known software PRAAT (Boersma and Weenink, 2011) as it contains a large set of primitives to perform phonetic analysis. Also, PRAAT is designed to efficiently handle multilayer annotations in terms of automatic generation, because of the scripting language, in terms of visualization, because of the built-in editors and drawing capabilites, and in terms of compatibility with external software, as the TextGrid format is widely supported. The Python implementations of the SOpS algorithm and of the syllabification algorithm are called from within PRAAT.

Since Prosomarker is designed to work on speech corpora, as soon as the user checks the desired options and presses the OK button in the main interface, the tool asks for the folder in which the audio (WAV) files can be found and, if any of the exporting options is set, it will ask for the folder in which to save results. In Figure A.2 a screenshot of the interface of the tool is shown.

Prosomarker can run both in automatic and semi-automatic mode: the user can select which steps of the annotation process he/she wishes to check manually and the tool will show the intermediate result waiting for confirmation before proceeding. It is also possible to go back to the target points manual positioning step after visualizing the automatically

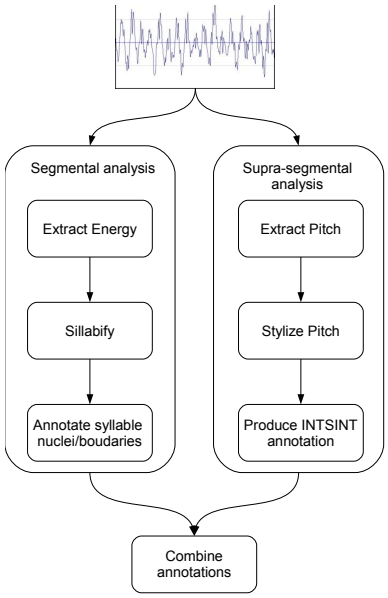


Figure A.1: The architecture of the Prosomarker tool

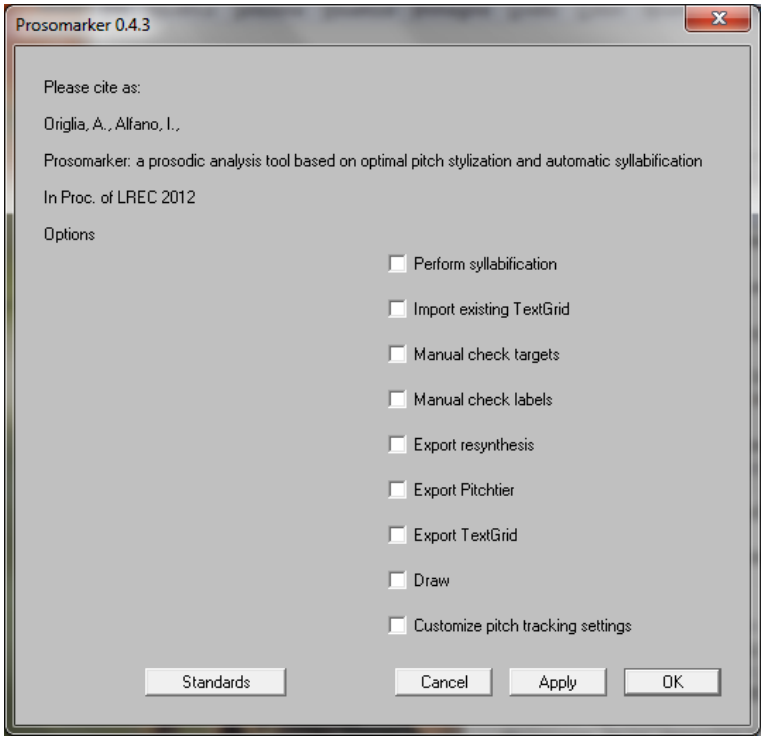


Figure A.2: The main interface of Prosomarker

assigned labels. This allows to check the positioning of the target points, to view and modify the labels and to introduce new tiers in the final TextGrid (for example, to introduce comments). Here I summarize all the different options the user can choose to customize how Prosomarker behaves:

- **Perform syllabification:** activates the segmental analysis process. Syllable nuclei and boundaries positions are automatically found and annotated in a separate tier.

- **Import existing TextGrid:** imports an existing annotation in TextGrid format and merges it with the requested output obtained from SOpS and from the syllabification algorithm. TextGrids to be imported must have the same name of the corresponding audio file.

- **Manual check targets:** activates the semi-automatic mode of Prosomarker. After performing the pitch stylization step, the tool will create a Manipulation object and open the corresponding PRAAT editor window in which the user can add target points, remove them or adjust their position.

- **Manual check labels:** activates the semi-automatic mode of Prosomarker. After performing the automatic annotation step, the tool opens a PRAAT editor window showing the waveform of the original sound file, its spectrum, pitch and intensity profile along with the produced annotations. Any operation available in PRAAT to manage TextGrids is available at this time. If the *Manual check targets* option was set, the possibility of going back to the target points adjustment step becomes available at run-time.

- **Export resynthesis:** instructs Prosomarker to generate a resynthesized version of the original sound file in which the stylized pitch curve is substituted to the original one by means of the PSOLA algorithm available in PRAAT. The resulting audio file is saved in the output directory set by the user.

- **Export PitchTier:** instructs Prosomarker to save the PitchTier object containing the target points used in the stylization process. If the user changes the target points, these changes will be saved. This can be useful to perform further analysis after running the tool.

- **Export TextGrid:** instructs Prosomarker to save the TextGrid containing the generated annotations. If the user modifies the TextGrid, the changes will be saved. This option is useful to transport data coming from the Prosomarker tool into other software supporting the TextGrid format.

- **Draw:** instructs Prosomarker to draw the original pitch curve along with its stylization and with the aligned TextGrid. In Figure A.3 we show an example of the automatic annotations Prosomarker produces generated with this option set.

- **Customize pitch tracking settings:** allows the user to customize pitch tracking settings used to obtain the pitch curve. The stylization process depends on the pitch curve extracted by PRAAT so these parameters influence the final SOpS curve.

Applications and future development

Prosomarker is an application designed to represent data coming from two algorithms dealing with different linguistic levels. The integrated visualization of these levels is proposed as a framework to provide researchers dealing with prosody an objective account of the occurrence of segmental and suprasegmental events along with their synchronization. Running in semi-automatic mode, the tool can be used both for fast data exploration and as a valid support to a prosodic analysis based on a phonetic approach: labels associated to target points not only provide a coherent description of global prosodic patterns, but they are also related with segmental events in such a way they can reveal linguistic regularities in the relationship between prosodic events and segmental string. Approaching speech from a perspective that tries to account for segmental and prosodic events simultaneously,

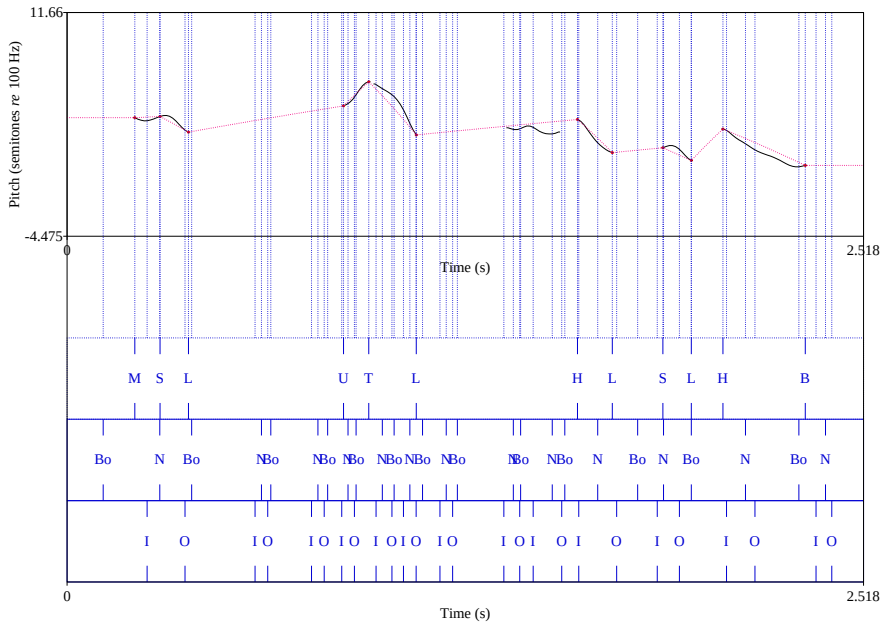


Figure A.3: An example of the annotation produced by Prosomarker. First tier: INTSINT labels. Second tier: syllable nuclei (N) and syllable boundaries (Bo). Third tier: extension of syllable nuclei from incipit (I) to offset (O).

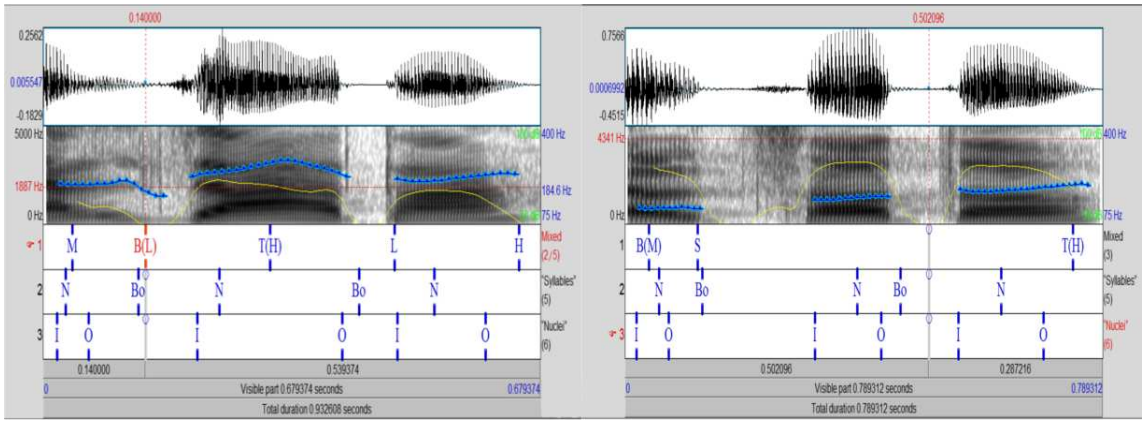


Figure A.4: The production of a native Italian speaker (on the left) compared with the production of a nonnative speaker (on the right). On the first tier: annotation labels; on the second tier: F0 differences of each target point compared with the previous one; on the third tier: duration increase; on the fourth tier: syllable nuclei (N) and syllable boundaries (Bo); on the fifth tier: extension of syllable nuclei from incipit (I) to offset (O)

but independently from each other, applies equally to quantitative and qualitative research strategies and offers the possibility to support a prosodic analysis considering different levels of detail within different theoretical frameworks. Depending on the specific research issues, Prosomarker can be used to analyze prosodic realizations related with linguistic modalities, pragmatic functions or emotion expressions, for instance.

The examples in Figure A.4 aim to show how Prosomarker can constitute a valid support in the interpretation of significant differences: they deal with the realizations of the same Italian question "E' alzata?" ("Is it standing up?") produced by a native Italian speaker and by a nonnative speaker in which we can observe how Prosomarker's annotations highlight differences in the comparison between L1 and L2 productions. The native realization is indeed characterized by a rising-falling contour (M - B(L) - T(H) - L - H) in which the maximum $F0$ value is aligned with the nucleus of the stressed vowel and with an important increase (almost 80 Hz), while the nonnative production presents a rising contour (B(M) - S - T(H)) in which $F0$ value increases progressively, reaching its maximum value at the end of the utterance.

In automatic mode, the tool can be used to rapidly process large sets of data for subsequent statistical analysis. In particular, the opportunity to produce an abstract representation of intonation involving different linguistic levels at the same time, but keeping them well separated is suitable to perform machine learning tasks.

Appendix B
Personality perception

Introduction

Social cognition has shown that people attribute, spontaneously and unconsciously, a wide range of socially relevant characteristics to others Uleman et al. (2008). Furthermore, the effect is so pervasive and ubiquitous that it takes place not only when people meet others in person, but also when others simply appear in audio and video recordings Reeves and Nass (1996). From a multimedia point of view, the main effect is that the perception of social and psychological phenomena taking place in the data influences significantly what we remember about the data we consume Dumais et al. (2003).

This work considers one aspect of this phenomenon, namely the spontaneous attribution of personality traits to unacquainted speakers. In particular, the article proposes an approach for *Automatic Personality Perception* (APP) based on *prosody*, the combination of (i) intonation, namely the combination of loudness, pitch, and speaking rate that characterizes the *way* someone speaks and (ii) voice quality, which reflects the way energy distribution across the frequency spectrum affects speech.

The main motivation for this choice is that the influence of both intonation and voice quality on personality perception has been extensively investigated in human sciences (e.g., see Scherer (1977)). Furthermore, domains like *Social Signal Processing* have shown that non-verbal behavioral cues (e.g. vocalizations, facial expressions, gestures, etc.) are a

reliable evidence for machine understanding of social, affective and psychological phenomena Vinciarelli et al. (2009).

To date, only a few approaches for APP have been proposed in the computing literature (see, e.g., Mairesse et al. (2007); Mohammadi and Vinciarelli (2012); Pianesi et al. (2008); Polzehl et al. (2010)). In contrast, the relationship between prosody and personality perception has been investigated for several decades in human sciences. The main findings can be summarized as follows: (i) high pitch variation tends to be perceived as higher competence and benevolence, and vice-versa Ray (1986), (ii) mean pitch tends to have negative correlation with respect to extraversion and dominance for females speakers, but positive correlation for male speakers Scherer (1977), and (iii) speaking rate tends to be positively correlated with perceived competence Ray (1986). In general, those findings suggest that prosody plays an important role in the way people perceive others.

To date, only a few approaches for APP have been proposed in the computing literature (see, e.g., Mairesse et al. (2007); Mohammadi et al. (2010); Mohammadi and Vinciarelli (2012); Pianesi et al. (2008); Polzehl et al. (2010)). In contrast, the relationship between prosody and personality perception has been investigated for several decades in human sciences. The main findings can be summarized as follows: (i) high pitch variation tends to be perceived as higher competence and benevolence, and vice-versa Ray (1986), (ii) mean pitch tends to have negative correlation with respect to extraversion and dominance for females speakers, but positive correlation for male speakers Scherer (1977), and (iii) speaking rate tends to be positively correlated with perceived competence Ray (1986); Smith et al. (1975). In general, those findings suggest that prosody plays an important role in the way people perceive others.

The experiments of this work, performed over the largest database of speakers assessed in terms of perceived personality traits, show that it is possible to predict the mutual position of two speakers in the personality space with up to 80% accuracy. The proposed approach is based on Ordinal Regression, which is the most suitable methodology to classify ordinally labeled data. To the best of our knowledge, this is the first work that goes beyond the simple prediction of traits attributed to speakers by predicting differences be-

tween individuals, in line with the cognitive processes behind personality perception Funder (2001).

APP can be beneficial for several technological domains, including the generation of synthetic voices capable of eliciting desired social perceptions (see Reeves and Nass (1996)), or the development of multimedia indexing approaches taking into account the way users perceive people portrayed in data Pantic and Vinciarelli (2009). More generally, APP can contribute to bridging both the social intelligence gap between people and machines and the semantic gap between the features and the content that people perceive in the data.

Personality: Model and Data

This section presents the personality model employed in this work and describes the data used in the experiments.

The "Big Five" Model

Personality is the latent construct accounting for "*individuals' characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms - hidden or not - behind those patterns*" Funder (2001). The *Big Five* (BF) personality model is the most commonly applied and accepted personality model Wiggins (1996) and proposes a personality representation based on five traits that have been shown to account for most of the individual differences:

- *Extraversion*: Active, Assertive, etc.
- *Agreeableness*: Appreciative, Kind, etc.
- *Conscientiousness*: Efficient, Organized, etc.
- *Neuroticism*: Anxious, Self-pitying, etc.
- *Openness*: Artistic, Curious, etc.

The *BF* model represents personalities in terms of five scores (one for each of the traits above) that can be obtained with appropriate assessment questionnaires. The scores

measure how well the adjectives accompanying the traits described a given individual. This work adopted the BFI-10 Rammstedt and John (2007), a short version (see Table B.1) of a longer questionnaire known as the *Big Five Inventory* (BFI) Rammstedt and John (2007). Each question in Table B.1 is associated to a Likert scale including five points ranging from "*Strongly disagree*" to "*Strongly agree*" and mapped into the interval $[-2, 2]$. The scores corresponding to each trait are obtained by simple numerical calculations performed over the answers to the questionnaire (see Rammstedt and John (2007) for more details). The main advantage of the BFI-10 is that it can be completed in less than a minute while still providing reliable results Rammstedt and John (2007).

The Data The experiments of this study were carried out over a corpus of 640 10 seconds long speech clips randomly extracted from the 96 news bulletins that *Radio Suisse Romande*, the French speaking Swiss national broadcast service, has broadcast during February 2005. There is one speaker per clip and the total number of unique speakers is 322. The personality assessment pool included 11 judges that have listened to each clip of the corpus and, immediately after listening, have filled the BFI-10 questionnaire. The judges have never met one another and have worked independently without being co-located (the assessment was performed via an online application). The judges have worked no more than 60 minutes per day (split into two 30 minutes sessions) to avoid tiredness effects. The clips have been presented to each judge in random order to cope with the reduction in attention observed towards the end of each session. The clips are in French and the 11 judges have signed a document stating that they do not speak or understand such language. This ensures that the content of the clips influences the personality assessment process only to a minor extent.

At the end of the assessment process, each clip is assigned five scores corresponding to the BFs. Each score is the average of the 11 scores assigned individually by the assessors. The average scores for each trait were then converted into N ordinal categories so that they represented a "degree" associated each personality trait. This was achieved by ordering the samples according to the corresponding score and then by splitting the resulting ranking

1	This person is reserved
2	This person is generally trusting
3	This person tends to be lazy
4	This person is relaxed, handles stress well
5	This person has few artistic interests
6	This person is outgoing, sociable
7	This person tends to find fault with others
8	This person does a thorough job
9	This person gets nervous easily
10	This person has an active imagination

Table B.1: The BFI-10 questionnaire used in the experiments (as proposed in Rammstedt and John (2007)).

into N equally sized groups.

The Approach

The proposed APP approach comprises three main steps: (i) extraction of short-term speech features by means of the method presented in Chapter 2, (ii) estimation of long-term statistical features, and (iii) mapping of those features into ordinal categories.

	$N = 3$			$N = 4$			$N = 5$			$N = 6$		
ρ	0%	50%	80%	0%	50%	80%	0%	50%	80%	0%	50%	80%
Ext.	78.6%	84.2%	88.8%	76.1%	79.5%	81.2%	75.0%	77.3%	76.8%	74.9%	76.9%	81.4%
Agr.	65.8%	69.0%	74.7%	63.6%	67.8%	76.9%	64.6%	67.5%	70.5%	64.1%	67.0%	70.6%
Con.	70.8%	74.8%	76.4%	69.4%	73.9%	81.7%	68.9%	73.6%	74.8%	68.2%	71.3%	75.6%
Neu.	72.0%	75.7%	77.8%	70.4%	74.2%	76.2 %	69.9%	73.4%	73.4%	69.0%	71.3%	69.4%
Ope.	63.9%	70.1%	69.3%	61.3%	64.7%	62.4%	61.6%	66.0%	69.6%	61.3%	65.6%	66.1%

Table B.2: Pairwise ranking results. The table reports the accuracy in predicting, for each trait, the speaker that has been scored higher by the assessors. The results were obtained for different numbers N of ordinal categories and different values ρ of rejection rate.

Statistical Features Estimation

At the end of the short-term feature extraction process, each nucleus is represented by a vector where each component corresponds to one of the features above. Statistics estimated over the feature values extracted from each nucleus individually are then used to represented a speech sample. In particular, the mean is computed for all features, the standard deviation is computed for nuclei and syllable length, pitch, energy, spectral slope, harmonicity, and spectral centroid, the entropy is estimated for nuclei and syllable length, pitch, energy, spectral slope, spectral centroid, and glissando likelihood. Mean and bandwidth of the first three formants are also extracted from each syllable nucleus The feature set is completed by the minimum of the pitch and the maximum energy. The total number of features is 35.

Ordinal Regression

Personality perception refers to the detection of phenotypic differences between individuals. Hence, the last step of the approach consists in automatically ranking people according to the personality traits attributed by human assessors. The most suitable method for such a purpose is *Ordinal Regression* (OR) McCullagh (1980). In OR, samples $\mathbf{x}_i$ are assigned to ordinal labels y_i belonging to the ordered set $C = (1, 2, \ldots, N)$. This work employs a linear probabilistic approach to OR as in McCullagh (1980).

Experiments and Results

The goal of APP is to rank people according to the personality traits attributed to them by human assessors. One way to evaluate the predictive power of an APP approach is to test its ability to rank correctly all possible pairs of test samples. In order to do so, consider a pair of test samples $\mathbf{x}_i, \mathbf{x}_j$ such that the corresponding labels for a given personality trait satisfy, say, $y_i > y_j$. The performance score is simply the average number of times that the APP approach predicts a label for $\mathbf{x}_i$ that is greater than the label predicted for $\mathbf{x}_j$

over the entire test set (the probability of being correct by chance is 50%). Given the probabilistic nature of the proposed APP approach, predicting the most likely ranking for a pair of test samples $\mathbf{x}_i$ and $\mathbf{x}_j$, with corresponding predictive probabilities $p(h_i|\mathbf{x}_i)$ and $p(h_j|\mathbf{x}_j)$, becomes

$$\arg\max_{(h_i,h_j)\in\mathcal{C}\times\mathcal{C}} \{p(h_i|\mathbf{x}_i)p(h_j|\mathbf{x}_j)\}\,, \tag{4.1}$$

with the constraint that $h_i \neq h_j$. In this application, the number of ordinal categories ranges from 3 to 6, and so the solution to eq. 4.1 is found by enumerating all possible rankings. Another advantage of taking a probabilistic approach, is that it is possible to reject the percentage ρ of samples where the ordinal regression approach is less confident about the prediction, as illustrated next.

In order to test the approach over the entire corpus while keeping a rigorous separation between training and test set, the experiments were performed using a K-fold validation procedure ($K = 15$) as follows. The corpus was split into K subsets of which $K - 1$ were used for training and one for testing. The folds were obtained randomly, but it was ensured that the same person did not appear in both training and test set. Performance were evaluated leaving one of the K folds out at each time and averaging the results obtained.

The second is to consider all N-tuples of test samples such that each element belongs to a *different* ordinal category (N is the number of ordinal categories), and to predict automatically the ordinal category each sample of the N-tuple belongs to. In the former case, the approach works correctly when it finds the sample that has actually been scored higher (the probability of being correct by chance is 0.5). In the latter case, the approach works correctly when each sample of the N-tuple has been assigned to its actual ordinal category (the probability of being correct by chance is $N!^{-1}$).

The first and simplest way of measuring the performance of the approach is to consider all pairs of samples in the test set and to measure the fraction of times that the approach ranks the two corresponding people correctly, i.e. the number of times that the person

with a higher score along a given trait is assigned a higher ordinal category. In this case, the probability of ranking correctly the samples is 50%.

The second way of measuring the performance is to measure the fraction of times that the approach ranks correctly N samples belonging to N different categories. Given a N-tuple of test samples $(x_1, \ldots, x_N)$, where the ordinal category of x_k is k and all samples belong to different categories, an Ordinal Regression approach finds the N-tuple of ordinal categories $H^* = (h_1^*, \ldots, h_N^*)$, where $h_i \neq h_j$ for $i \neq j$, that satisfies the following equation:

$$H^* = \arg\max_{H \in \mathcal{H}} \prod_{k=1}^{N} \pi_{h_k}(x_k) \tag{4.2}$$

where $\mathcal{H}$ is the set of all possible N-tuples of ordinal categories where $h_i \neq h_j$ for $i \neq j$.

When H^* is such that h_k^* is the actual category of x_k for all k values, it means that the approach has ranked correctly the samples of the test N-tuple. Therefore, the performance of the approach can be measured by considering all test N-tuples where each sample belongs to a different class and by considering the fraction of times that all N-tuple samples are assigned to the correct ordinal category. The probability ranking correctly the samples of the N-tuple is $(N!)^{-1}$.

Results

Table B.2 reports the results obtained using models with $N = 3, 4, 5$ and 6 (and $\rho = 0\%, 50\%$ and 80%.). The higher the number of ordinal categories N, the higher the resolution at which it is possible to discriminate between people. The performance difference with respect to chance is always statistically significant with p-value $p < 5\%$. The results suggest that the approach is robust with respect to the number of ordinal categories as no major performance losses are observed when going from $N = 3$ to $N = 6$. The influence of ρ depends on the particular trait, but the general trend is of an increase by roughly 5% when going from no rejection to $\rho = 50\%$, and by another 5% when further increasing ρ to 80%.

According to the indications of the psychological literature, the prediction of Extraversion and Conscientiousness achieves, on average, higher performance. The reason is that, from a cognitive point of view, these are the two most accessible traits Funder (2001). In contrast, the good performance on Neuroticism seems to be a peculiarity of the dataset and it probably depends on the polarization of the assessments (many subjects tend to be assigned to the extremes of the scale). From an acoustic point of view, the correlation of every feature with every trait is significant with $p < 5\%$. In this respect, the results of this work confirm the psychological findings mentioned in the Section 4.4.

	$N = 3$ ($p = 16.7\%$)			$N = 4$ ($p = 4.2\%$)			$N = 5$ ($p = 0.8\%$)			$N = 6$ ($p = 0.1\%$)		
ρ	0%	50%	80%	0%	50%	80%	0%	50%	80%	0%	50%	80%
Ext.	50.3%	55.9%	65.7%	21.3%	26.5%	48.2%	8.8%	15.7%	14.7%	3.1%	6.3%	23.3%
Agr.	30.8%	32.1%	41.2%	10.4%	13.8%	25.1%	3.5%	5.0%	21.8%	0.8%	2.9%	12.0%
Con.	39.3%	45.2%	45.9%	15.8%	21.1%	36.6%	5.8%	10.3%	19.1%	1.9%	5.6%	7.8%
Neu.	38.7%	41.3%	41.6%	15.0%	16.1%	30.6%	5.4%	5.0%	28.6%	1.2%	2.4%	12.7%
Ope.	29.7%	31.8%	46.0%	9.6%	13.4%	21.8%	2.8%	3.2%	21.9%	0.6%	1.8%	11.7%

Table B.3: Sequence ranking results. The table reports the accuracy in ranking N-tuples including as many samples as the ordinal categories. The results were obtained for different numbers N of ordinal categories and different values ρ of rejection rate. The value p is the chance performance.

Sequence Ranking

The experiments are similar to pairwise ranking, but in this case the goal is to rank correctly N test samples, each belonging to a different ordinal category. Equation (4.1) can easily be extended to the case of a N-tuple $(h_1, \ldots, h_N) \in C^N$, where $h_i \neq h_j$ for $i \neq j$. Table B.3 reports the results for the same values of N and ρ used for pairwise ranking. In this case as well, the difference with respect to chance performance is statistically significant for all reported values . The performance is satisfactory only for high values of ρ and low values of N. However, it must be noted that a sequence is considered wrong even if only one of the elements is in the wrong order. The high performance in pairwise ranking suggests that, in most cases, most of the elements of an N-tuple are still in the right order.

Conclusions

The key elements of the proposed approach are (i) the use of features extracted from intonation and voice quality, (ii) the use of a probabilistic approach to map such features into the personality space, and (iii) a thorough evaluation based on the largest database of personality assessments from radio broadcasts available in the literature. These results, published in Mohammadi et al. (2012), show that it is possible to automatically rank people with different degrees of personality traits with an accuracy around 80%.

List of Figures

List of Tables

Bibliography

A., A. L. and A., H. B. (2002). Zipf's law and the internet. *Glottometrics*, pages 143–150.

Abete, G., Cutugno, C., Ludusan, B., and Origlia, A. (2010). Pitch behavior detection for automatic prominence recognition. In *Proc. of Speech Prosody*.

Arnold, M. B. (1960). *Emotion and personality*. Cambridge University Press.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, pages 46–63.

Avanzi, M., Lacheret-Dujour, A., and Victorri, B. (2008). Analor. a tool for semi-automatic annotation of french prosodic structure. In *Proc. of Speech Prosody*, pages 119–122.

Avanzi, M., Lacheret-Dujour, A., and Victorri, B. (2010). A corpus based learning method for prominence detection inspontaneous speech. In *Proc. of Speech Prosody*.

Barakova, E. I. and Lourens, T. (2010). Expressing and interpreting emotional movements in social games with robots. *Personal and Ubiquitous Computing*, 14(5):457–467.

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., and Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? In *Proc. of ICPhS*, pages 2693–2696.

Batliner, A., Seppi, D., Steidl, S., and Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human–Computer Interaction - Special Issue on emotion-aware natural interaction*.

Batliner, A., Steidl, S., and Elmar, N. (2011). Associating children's non-verbal and verbal behaviour: Body movements, emotions, and laughter in a human-robot interaction. In *Proc. of ICASSP*, pages 5828–5831.

Becker, C., Kopp, S., and Wachsmuth, I. (2004). *Simulating the emotion dynamics of a multimodal conversational agent*, pages 154–165. Springer.

Bevacqua, E., Prepin, K., Niewiadomski, R., de Sevin, E., and Pelachaud, C. (2010). Greta: Towards an interactive conversational virtual companion. In *Artificial companions in society: perspectives on the present and future*, pages 1–17.

Bloomfield, L. (1933). *Language*. Reinhart and Wiston.

Boersma, P. and Weenink, D. (2011). Praat: doing phonetics by computer [computer program]. version 5.2.40.

Bolinger, D. (1958). A theory of pitch accent in English. *Word*, pages 109–149.

Breitenstein, C., Van Lancker, D., and Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a german and an american sample. *Cognition and Emotion*, pages 57–79.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, pages 3–15.

Burattini, E. and Rossi, S. (2010). Periodic activations of behaviours and emotional adaptation in behaviour-based robotics. *Connection Science*, 22:297–213.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of german emotional speech. In *Proc. of Interspeech, Lisbon*, page 1517–1520.

Campione, E., Hirst, D., and Véronis, J. (2000). Stylisation and symbolic coding of f0: comparison of five models. In Botinis, A., editor, *Intonation: analysis, modeling and technology*, pages 185–208, Dordrecht. Kluwer.

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., and Paouzaiou, A. K. K. (2006). Modeling naturalistic affective states via facial and vocal expression recognition. In *Proc. of ICMI*, pages 146–154.

Carnahan, J. and Sinha, R. (2001). Nature's algorithms [genetic algorithms]. *IEEE Potentials*, 20(2):21–24.

Chang, C. and Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on intelligent systems and technology*, 2:27:1 – 27:27.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M. (2000). Feeltrace: An instrument for recording perceived emotion in real time. In *Proc. of ISCA Workshop on Speech & Emotion*, pages 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human–computer interaction. *IEEE Signal Processing Magazine*, 18:33–80.

D'Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, 9(3):257–288.

Darwin, C. (1872). *The expression of emotions in man and animals*. John Murray, Albemarle street, London.

De Saussure, F. (1967). *Course de linguistique generale*. Laterza, payot. italian edition by t. de mauro edition.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ΔC. In Karnowski, P. and Szigeti, I., editors, *Language and language processing*, pages 231–241. Peter Lang: Frankfurt am Main.

Dellwo, V. and Wagner, P. (2003). Relationships between speech rate and rhythm. In *Proc. of the ICPhS*, pages 471–474.

D'Imperio, M. and House, D. (1997). Perception of questions and statements in neapolitan italian. In *Proc. of EUROSPEECH*, pages 22–25.

Drioli, C., Tisato, G., Cosi, P., and Tesser, F. (2003). Emotions and voice quality: Experiments with sinusoidal modeling. In *Proc. of VOQUAL*, pages 127–132.

Dumais, S., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R., and Robbins, D. C. (2003). Stuff i've seen: a system for personal information retrieval and re-use. In *Proc. of ACM Intl. Conf. on Research and Development in Information Retrieval*, pages 72–79.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, pages 169–200.

Eriksson, A., Grabe, E., and Traunmueller, H. (2002). Perception of syllable prominence by listeners with and without competence in the tested language. In *Proc. of Speech Prosody*, pages 275–278.

Espinosa, H. P., Garcia, C. A. R., and Pineda, L. V. (2010). Features selection for primitives estimation on emotional speech. In *Proc. of IEEE ICASSP*, pages 5138–5141.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Fernandez, R. and Picard, R. (2011). Recognizing affect from speech prosody using hierarchical graphical models. *Speech Communication*, pages 1088–1103.

Feth, L. L. (1972). Combinations of amplitude and frequency differences in auditory discrimination. *Acustica*, 26:67–77.

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two dimensional. *Psychological Science*, 18:1050–1057.

Fragopanagos, N. and Taylor, J. G. (2005). Special issue: Emotion recognition in human-computer interaction. *Neural Networks*, 18:389—405.

Frederick, S. andLoewenstein, G. (1999). Hedonic adaptation. In Diener, E., Schwartz, N., and Kahneman, D., editors, *Hedonic psychology: Scienti?c approaches to enjoyment, suffering, and wellbeing*, pages 302—-329. New York: Russell Sage Foundation Press.

Funder, D. (2001). Personality. *Annual Reviews of Psychology*, 52:197–221.

Gharavian, D., Sheikhan, M., and Ashoftedel, F. (2012). Emotion recognition improvement using normalized formant supplementary features by hybrid of dtw-mlp-gmm model. *Neural Computing and Applications*, pages 1–11.

Ghosh, P. K. and Narayanan, S. (2009). Pitch Contour Stylization Using an Optimal Piecewise Polynomial Approximation. *IEEE Signal Processing Letters*, 16(9):810–813.

Gordon, S. M., Kawamura, K., and Wilkes, D. M. (2010). Neuromorphically inspired appraisal-based decision making in a cognitive robot. *IEEE Transaction on autonomous mental development*, 2.

Goudbeek, M., Goldman, J. P., and Scherer, K. R. (2009). Emotion dimensions and formant position. In *Proc. of InterSpeech*, pages 1575–1578.

Graham, G. H., Unruh, J., and Jennings, P. (1991). The impact of nonverbal communication in organizations: A survey of perceptions. *Journal of Business Communication*, 28:45–62.

Greenberg, S. and Kingsbury, B. E. (1997). The modulation spectrogram: In pursuit of an invariant representation of speech. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, pages 1647–1650.

Gregory, M. L. (2004). Using conditional random fields to predict pitch accents in conversational speech. In *Proc. of ACL [Online]*.

Grimm, M. and Kroschel, K. (2005). Emotion estimation in speech using a 3D emotion space concept. In *Proc. of IEEE Automatic Speech Recognition & Understanding Workshop*, pages 381–385.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The vera am mittag german audio-visual emotional speech database. In *Proc. of ICME*, pages 865–868.

Gunes, H., Schuller, B., Pantic, M., and Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. of FG*, pages 827–834.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, Hamilton, New Zealand.

Hirst, D., Di Cristo, A., and Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In Horne, M., editor, *Prosody: Theory and Experiment Studies*, Dordrecht. Kluwer.

Hirst, D. and Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phon?tique d'Aix-en-Provence*, 15:75–85.

House, D. (1990). *Tonal perception in speech*. Lund University Press, Lund.

House, D. (1995). Perception of prepausal tonal contours: implications for automatic stylization of intonation. In *Proc. of Eurospeech*, pages 949–952.

House, D. (1996). Differential perception of tonal contours through the syllable. In *Proc. of ICSLP*, pages 2048–2051.

Ioannou, S., Kessous, L., Caridakis, G., Karpouzis, K., Aharonson, V., and Kollias, S. (2006). Adaptive on-line neural network retraining for real life multimodal emotion recognition. In *Proc. of ICANN*, pages 81–92.

James, W. (1884). What are emotions? *Mind*, 9:188–205.

Jensen, C. (2003). Perception of prominence in standard british english. In *Proc. of ICPhS*.

Jespersen, O. (1920). *Lehrbuch der Phonetic*. B.G. Teubner, Leipzig e Berlin.

Jia, J., Zhang, S., Meng, F., Wang, Y., and Cai, L. (2011). Emotional audio-visual speech synthesis based on PAD. *IEEE Transactions on Audio, Speech & Language Processing*, pages 570–582.

Jittiwarangkul, N., Jitapunkul, S., Luksaneeyanavin, S., Ahkuputra, V., and Wutiwiwatchai, C. (1998). Thai syllable segmentation for connected speech based on energy. In *Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS'98)*, pages 169–172.

Jones, R. J., Downey, S., and S., M. J. (1997). Continuous speech recognition using syllables. In *Proceedings of Eurospeech*, pages 1171–1174.

Jürgens, R., Hammerschmidt, K., and Fischer, J. (2011). Authentic and play-acted vocal emotion expressions reveal acoustic differences. *Frontiers in Psychology*, 2.

Kao, Y. and Lee, L. (2006). Feature analysis for emotion recognition from mandarin speech considering the special characteristics of chinese language. In *Proc. of Interspeech*, pages 1814–1817.

Kim, H.-R. and Kwon, D.-S. (2010). Computational model of emotion generation for human-robot interaction based on the cognitive appraisal theory. *Journal of Intelligent and Robotic Systems*, 60:263–283.

Kirpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by simulated annealing. *Science*, pages 671–680.

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *Journal of the Acoustical Society of America*, 53:8–16.

Kozhevnikov, V. and Chistovich, L. (1966). *Motor phonetics. Tech. rep.* U.S. Department of Commerce. Joint Publication Research Service, Washington DC.

Kuremoto, T., Yamane, T., Feng, L., Kobayashi, K., and Obayashi, M. (2011). A human-machine interaction system: A voice command learning system using pl-g-som. In *Proc. of MASS*, pages 1–4.

Ladd, R. D. (1996). *Intonational phonology*. Cambridge University Press.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pages 282–289.

Ledoux, J. (1998). *The emotional brain: the mysterious underpinnings of emotional life*. Simon & Schuster.

Leventhal, H. and Scherer, K. (1987). The relationship of emotion to cognition: a functional approach to semantic controversy. *Cognition and Emotion*, 1(1):3–28.

Ludusan, B. (2010). *Beyond short units in speech recognition: a syllable-centric and prominence-based approach*. PhD thesis, Universitá di Napoli "Federico II".

Ludusan, B., Origlia, A., and Cutugno, C. (2011). On the use of the rhythmogram for automatic syllabic prominence annotation. In *Proc. of Interspeech*, pages 2413–2416.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

Maiwald, D. (1967). Ein funktionschema des gehörs zur beschreibung der erkennbarkeit kleiner frequenz-und-amplitudeänderungen. *Acustica*, 18:81–93.

McCullagh, P. (1980). Regression models for ordinal data. *Journal Royal Statistical Society B*, 42:109–142.

McKeown, G., Valstar, M. F., Cowie, R., and Pantic, M. (2010). The semaine corpus of emotionally coloured character interactions. In *Proc. of ICME*, pages 1079–1084.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology: Developmental, Learning, Personality, Social*, 14:261–292.

Mermelstein, D. (1975). Automatic segmentation of speech into syllabic units. *Journal of Acoustical Society of America*, 54(4):880–883.

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In *Proc. of Speech Prosody*.

Mertens, P. (2006). A predictive approach to the analysis of intonation in discourse in french. In Kawaguchi, Y., Fonagy, I., and Moriguchi, T., editors, *Prosody and Syntax*, pages 64–101, Amsterdam. John Benjamins.

Millward, C. M. (1996). *A biography of the English language*. Harcourt Brace.

Minsky, M. L. (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster.

Mohammadi, G., Origlia, A., Filippone, M., and Vinciarelli, A. (2012). From speech to personality: Mapping voice quality and intonation into personality differences. In *Proc. of ACM Multimedia*, pages 789–792.

Mohammadi, G. and Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3):273–284.

Mohammadi, G., Vinciarelli, A., and Mortillaro, M. (2010). The voice of personality: Mapping nonverbal vocal behavior into trait attributions. *In Proc. of Intl. Wks. on Social Signal Processing*, pages 17–20.

Morency, L., Quattoni, A., and Darrell, T. (2007). Latent-dynamic discriminative models for continuous gesture recognition. In *Proc. of CVPR*, pages 1–8.

Nagarajan, T., Murthy, H. A., and Hegde, R. M. (2003). Segmentation of speech into syllable-like units. In *Proceedings of Eurospeech 2003*, pages 2893–2896.

Nygaard, R. and Haugland, D. (1998). Compressing ECG signals by piecewise polynomial approximation. In *Proc. of ICASSP*, volume 3, pages 1809–1812.

Oliver, D. (2005). Deriving pitch accent classes using automatic f_0 stylization and unsupervised clustering techniques. In *Proc. of Second Baltic Conference on Human Language Technologies*, pages 161–166.

Origlia, A., Abete, G., Cutugno, C., Alfano, I., Savy, R., and Ludusan, B. (2011). A divide et impera algorithm for optimal pitch stylization. In *Proc. of Interspeech*, pages 1993–1996.

Origlia, A., Abete, G., and Cutugno, F. (2013). A dynamic tonal perception model for optimal pitch stylization. *Computer Speech and Language*, 27:190–208.

Origlia, A. and Alfano, I. (2012). Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification. In *Proc. of LREC-2012*, pages 997–1002.

Pantic, M. and Vinciarelli, A. (2009). Implicit human-centered tagging. *IEEE Signal Processing Magazine*, 26(6):173–180.

Patel, A. D. (2005). The relationship of music to the melody of speech and to syntactic processing disorders in aphasia. *Annals of the New York Academy of Sciences*, 1060:59–70.

Patel, S., Scherer, K. R., Bjorkner, E., and Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, pages 93–98.

Petek, B., Andersen, O., and Dalsgaard, P. (1996). On the robust automatic segmentation of spontaneous speech. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP)*, pages 913–916.

Petrillo, M. and Cutugno, F. (2003). A syllable segmentation algorithm for english and italian. In *Proc. of Eurospeech*, pages 2913–2916.

Pfitzinger, H., Burger, S., and Heid, S. (1996). Syllable detection in read and spontaneous speech. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP)*, pages 1261–1264.

Pianesi, F., Mana, N., and Cappelletti, A. (2008). Multimodal recognition of personality traits in social interactions. In *In Proc. of the Intl. Conf. on Multimodal Interfaces*, pages 53–60.

Picard, R. (1997). *Affective computing*. MIT Press.

Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology.

Pollack, I. (1968). Detection of rate of change of auditory frequency. *Journal of experimental psychology*, 77:535–541.

Polzehl, T., Moller, S., and Metze, F. (2010). Automatically assessing personality from speech. *In Proc. of IEEE Intl. Conf. on Semantic Computing*, pages 134–140.

Poole, H. (1985). *Theories of the middle range*. Ablex Pub. Corp., Norwood, N.J.

Prasad, V. K., Nagarajan, T., and Murthy, H. A. (2004). Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communication*, pages 429–446.

Rammstedt, B. and John, O. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212.

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, pages 265–292.

Ray, G. B. (1986). Vocally cued personality prototypes: An implicit personality theory approach. *Journal of Communication Monographs*, 53(3):266–276.

Reeves, B. and Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Reichl, W. and Ruske, G. (1993). Syllable segmentation of continuous speech with artificial neural networks. In *Proceedings of Eurospeech93, 3rd European Conference on Speech Communication and Technology*, pages 1771–1774.

Roach, P. (2000). *English Phonetics and Phonology. A Practical Course*. CUP.

Roseman, I. J. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion*, 10(3):241–278.

Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica*, 23:1–33.

Rossi, M. (1972). Interactions of intensity glides and frequency glissandos. *Language and Speech*, 21:384–394.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality & Social Psychology*, 39:1161–1178.

Savy, R. and Cutugno, F. (2009). Clips: diatopic, diamesic and diaphasic variations of spoken italian. In *Proc. of Corpus Linguistics Conference*.

Sawusch, J. R. (2005). Acoustic analysis and synthesis of speech. In Pisoni, D. B. and Remez, R. E., editors, *The handbook of speech perception*, pages 7–27, Malden (MA) et al. Blackwell.

Scherer, K. R. (1977). Effect of stress on fundamental frequency of the voice. *in Journal of Acoustical Society of America*, 62(S1):25–26.

Scherer, K. R., Johnstone, T., and Klasmeyer, G. (2003). Vocal expression of emotions. In Davidson, R. J., Scherer, K. R., and Goldsmith, H. H., editors, *Handbook of Affective Sciences*, pages 433–456. Oxford University Press.

Scherer, K. R., Shorr, A., and Johnstone, T. (2001). *Appraisal processes in emotion: theory, methods, research*. Oxford University Press.

Schouten, H. E. M. (1985). Identification and discrimination of sweep tones. *Perception and psychophysics*, 37:369–376.

Schuller, B., Batliner, A., Steidl, S., and D., S. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, pages 1062–1087.

Schuller, B., Vlasenko, B., Arsic, D., Rigoll, G., and Wendemuth, A. (2008). Combining speech recognition and acoustic word emotion models for robust textindependent emotion recognition. In *Proc. of IEEE ICME*, pages 1333—1336.

Seppi, D., Batliner, A., Steidl, S., Schuller, B., and Nöth, E. (2010). Word accent and emotion. In *Proc. of Speech Prosody*.

Sergeant, R. L. and Harris, J. D. (1962). Sensitivity to unidirectional frequency modulation. *Journal of the Acoustical Society of America*, 34:1625–1628.

Shamsuddin, S., Yussof, H., Ismail, L., Hanapiah, F., Mohamed, S., Piah, H., and Ismarrubie Zahari, N. (2012). Initial response of autistic children in human-robot interaction therapy with humanoid robot nao. In *Proc. of CSPA*, pages 188–193.

Shastri, L., Chang, S., and Greenberg, S. (1999). Syllable detection and segmentation using temporal flow neural networks. In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, pages 1721–1724.

Siemund, R., Höge, H., Kunzmann, S., and Marasek, K. (2000). Speecon-speech data for consumer devices. In *Proc. of LREC*, pages 883–886.

Silipo, R. and Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous english discourse. In *Proc. of ICPS*.

Silipo, R. and Greenberg, S. (2000). Automatic transcription of prosodic stress for spontaneous english discourse. In *Proc. of ICPS*.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). Tobi: a standard for labeling english prosody. In *Proc. of ICSLP*, pages 13–16.

Smith, B. L., Brown, B. L., Strong, W. J., and Rencher, A. C. (1975). Effect of speech rate on personality perception. *Journal of Language and Speech*, 18:146–152.

Sridhar, V. K. R., Nenkova, A., Narayanan, S., and Jurafsky, D. (2008). Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proc. of Speech Prosody [Online]*, pages 453–456.

Stetson, R. (1951). *Motor Phonetics*. North Holland, Amsterdam.

Tamburini, F. (2006). Reliable prominence identification in english spontaneous speech. In *Proc. of Speech Prosody*.

Tamburini, F. and Wagner, P. (2007). On automatic prominence detection for german. In *Proc. of Interspeech*, pages 1809–1812.

Tao, J. and Tieniu, T. (2005). Affective computing: A review. In *Affective Computing and Intelligent Interaction*, pages 981—995. Springer.

Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107:1697–1714.

Terken, J. (1991). Fundamental frequency and perceived prominence. *Journal of the Acoustical Society of America*, pages 1768–1776.

t'Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America*, 69:811–821.

t'Hart, J., Collier, R., and Cohen, A. (1990). *A Perceptual Study of Intonation: An Experimental-Phonetic Approach*. Cambridge University Press, Cambridge.

Uleman, J. S., Saribay, S. A., and Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Reviews of Psychology*, 59:329–360.

Vanderslice, R. and Ladefoged, P. (1972). Binary suprasegmental features and transformational word-accentuation rules. *Language*, 48:819–836.

Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer Verlag.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag.

Vassière, J. (2005). Perception of intonation. In Pisoni, D. B. and Remez, R. E., editors, *The handbook of speech perception*, pages 236–263, Malden (MA) et al. Blackwell.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759.

Vlasenko, B., Philippou-Hübner, D., Prylipko, D., Böck, R., Siegert, I., and Wendemuth, A. (2011). Vowel formants analysis allows straightforward detection of high arousal emotions. In *Proc. of ICME*, pages 4230–4235.

Vogt, T. and Andre, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proc. of ICME*, pages 474–477.

Von Bekesy, G. (1960). *Experiments in hearing*. McGraw-Hill.

Wang, D. and Narayanan, S. (2005). Piecewise linear stylization of pitch via wavelet analysis. In *Proc. of the European Conference on Speech Communication and Technology*, pages 1–4.

Watrous, R. L. (1993). Gradsim: a connectionist network simulator using gradient optimization techniques. Technical report, Siemens Corporate Research Inc, Princeton, New Jersey.

Wiggins, J., editor (1996). *The Five-Factor Model of Personality*. Guildfor Press.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., and Cowie, R. (2008). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. of Interspeech*, pages 597–600.

Wu, S., Falk, T. H., and Chan, W. (2009). Automatic recognition of speech emotion using long-term spectro-temporal features. In *Proc. of ICDSP*, pages 1–6.

Wu, S., Falk, T. H., and Chan, W. (2010). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53:768–785.

Wu, S., Falk, T. H., and Chan, W. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53:768–785.

Wu, S., Shire, M. L., Greenberg, S., and Morgan, N. (1997). Integrating syllable boundary information into speech recognition. In *Proceedings of ICASSP*, pages 987–990.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An Introduction to Human Ecology*. Addison-Wesley Press.

Zwicker, E. (1962). Direct comparisons between the sensations produced by frequency modulation and amplitude modulation. *Journal of the Acoustical Society of America*, 34:1425–1430.