Exploring interactions between music and language during the early development of music cognition. A computational modelling approach.

Inês Salselas

Tesi Doctoral UPF / 2013

Director de la tesi:
Dr Xavier Serra
Dept. of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona, Spain

This research was performed at the Music Technology Group of the
Universitat Pompeu Fabra in Barcelona, Spain.
This research was funded by Fundació Barcelona Media.

"A música é tamanha
cabe em qualquer medida."

Music is all-embracing,
it fits any measure.

Sérgio Godinho

Acknowledgments

It was a great opportunity to be part of the MTG. It changed me and opened a new window for looking towards science and the world. I owe this to Xavier Serra. He also gave me the chance to be part of the EmCAP project, which provided a motivational context for this research. Without these opportunities, this thesis would never have happened.

I am truly grateful to Perfecto Herrera, who supported me since the very beginning. Ideas do not come to life in isolation and Perfecto knows how to make them flourish, give them shape, guide them ahead and enable them to be something greater.

I would like to thank everyone from the MTG who made my experience there very pleasant. Specially to my office colleagues and lunch palls Amaury Hazan, Ferdinand Fuhrmann, Dmitry Bogdanov, Gerard Roma, Graham Coleman, Hendrik Purwins, Martín Haro, Ricard Marxer, Piotr Holonowicz. With them, music was for sure the most amazing and intriguing subject to study in the world.

I would like to acknowledge the generosity of the caregivers who were willing to open and record their privacy and give a great contribute to this research. I am very grateful to them all.

During this time, I have been in contact with researchers whose conversations and tips influenced this work. They are São Luís Castro and Luís Gustavo Martins. I am very grateful for their inspiration and constant availability to help me.

Denis Mareschal opened up the opportunity for a fruitful research stay at the Centre for Brain and Cognitive Development at Birkbeck College. During this period I had the chance to be in contact with excellent researchers in a stimulating environment. Denis was also responsible for the inspiration for the last part of this research. For that I am very grateful.

I would also like to thank to Daniela Coimbra, Mário Azevedo and Octávio Inacio from ESMAE, IPP, for their kindness in allowing me to work in their lab during the writing period of the thesis.

Above all, I would like to acknowledge my family and friends. I thank to Manel and Ana for their love. I thank to my mother for her unconditional support and for expecting nothing but excellence from me. I thank to my

father for having taught me the value of knowledge and that life without love and poetry is not complete. I thank to Luís for making my everyday routine lighter and brighter.

Abstract

This dissertation concerns the computational modelling of early life development of music perception and cognition. Experimental psychology and neuroscience show results that suggest that the development of musical representations in infancy, whether concerning pitch or rhythm features, depend on exposure both to music and language. Early musical and linguistic skills seem to be, therefore, tangled in ways we are yet to characterize. In parallel, computational modelling has produced powerful frameworks for the study of learning and development. The use of these models for studying the development of music information perception and cognition, connecting music and language still remains to be explored.

This way, we propose to produce computational solutions suitable for studying factors that contribute to shape our cognitive structure, building our predispositions that allow us to enjoy and make sense of music. We will also adopt a comparative approach to the study of early development of musical predispositions that involves both music and language, searching for possible interactions and correlations.

We first address pitch representation (absolute vs relative) and its relations with development. Simulations have allowed us to observe a parallel between learning and the type of pitch information being used, where the type of encoding influenced the ability of the model to perform a discrimination task correctly. Next, we have performed a prosodic characterization of infant-directed speech and singing by comparing rhythmic and melodic patterning in two Portuguese (European and Brazilian) variants. In the computational experiments, rhythm related descriptors exhibited a strong predictive ability for both speech and singing language variants' discrimination tasks, presenting different rhythmic patterning for each variant. This reveals that the prosody of the surrounding sonic environment of an infant is a source of rich information and rhythm as a key element for characterizing the prosody from language and songs from each culture. Finally, we built a computational model based on temporal information processing and representation for exploring how the temporal prosodic patterns of a specific culture influence the development of rhythmic representations and predispositions. The simulations show that exposure to the surrounding sound environment influences the development of temporal representations and that the structure of the exposure environment, specifically the lack of maternal songs, has an impact on how the model organizes its internal representations.

We conclude that there is a reciprocal influence between music and language. The exposure to the structure of the sonic background influences the shaping of our cognitive structure, which supports our understanding of musical experience. Among the sonic background, language's structure has a predominant role in biasing the building of musical predispositions and representations.

Contents

List of Tables

List of Figures

1

Introduction

1.1 Motivation

In this research, we address, from a computational perspective, the way the exposure to specific registers of music and language interplay in early stages of development. Our methodology and computer experiments aim to shed light on the ways speech prosodic aspects, which are specific of the native language, can be reflected on or enhanced by implicit musical enculturation.

Music is part of human existence. Throughout human evolution, in every culture, music has been present. Music, in its earliest manifestations, was central in group activities, functioning as a means of establishing behavioural coherence in groups of people in contexts such as ritual and religious ceremonies and occasions calling for military arousal (Roederer, 1984). Music was also functional in regulating emotions in mother-infant interaction (Masataka, 2009).

Nowadays, in modern western societies, music is part of the fabric of everyday life. As in the past, music still remains embedded in collective celebrations such as sports events or weddings, and in contexts where it conveys emotional states, as in movies and advertisement. Our cognitive architecture allows us to compose music, to perform it and listen to it and, while in its turn it may be modulated and changed by these activities.

Still, with the growing focus in the individual, a different dimension of music experience emerges: an awareness of music as an individual's conscious experimentation, with self-regulation functions, characterized by differentiated sensitivities. Every human being has the capacity to understand and use music, but each one's music experience acquires a subjective meaning and brings forth a unique feeling. But how do we form these individual preferences? Are they innate? Or, alternatively, are they brought about by experience?

Early experience has a fundamental role in brain development. During infancy the brain develops rapidly, experiencing peak synaptic activity and forming neural networks. In this critical period, developmental processes are especially sensitive to environmental input, and the acquisition of adult

level abilities in specific areas is dependent on the surrounding stimuli or the lack of them (Patel, 2008). Exposure to sonic information, along with genetic influence, is determinant for the strengthening of neural communication paths, synaptic formation and organization. So, could it be that we build our predispositions through development? Does the information present in the sonic environment of an infant influences its musical aesthetics and bias its preferences?

Among the auditory information to which infants are exposed, the most salient are speech and singing sounds. Parents and caregivers, across cultures, languages and musical systems, use a distinctive register for singing and speaking to their infants (Papoušek & Papoušek, 1991; Trehub et al., 1993). The distinctive modifications that are characteristic of this register attract the infant's attention and may be used by adults to regulate infant's mood, playing an important role in conveying a range of communicative intentions to pre-verbal infants (Rock et al., 1999; Fernald, 1993).

From the perspective of a pre-verbal infant, music and speech may be not as differentiated as they are for older children and adults. Both music and language may be perceived as sound sequences that unfold in time, containing elements such as frequency, intensity, and duration. Thus, from the perspective of a pre-linguistic infant, who must learn about each system before understanding its communicative intent, music and language may be very similar. Consequently, at an early developmental stage, infants may use a single mechanism underlying both learning domains (McMullen & Saffran, 2004). So, could it be that mother tongue has a paramount role in the process of shaping our musical predispositions? Does this bring a cultural imprint into our cognitive architecture?

However, how can we approach this complex problem? If each human being develops her own predispositions driven by her unique experience, how can we avoid a "mythopoeic explanation"[1] (Cook, 1992) to surpass the bounds of the intrinsic individuality of this phenomenon? Can we find answers to our problem by focusing on the building up of the mechanisms that underlie this process? If we study the mechanisms behind the shaping of individual sensitivity, will we be able to understand subjective experiences of music?

Computational tools propose a formal and explicit method for specifying how information is processed in the brain (Shultz, 2003). This way, computer models bring out the conditions to go beyond a descriptive approach that accounts on a superficial perspective of human behaviours that may be considered an emergent result of inner working processes and mechanisms. Hence, the use of computational tools aims to explain the intrinsic paramet-

[1] In the sense that the scientific method approach to music must be rejected since music is a cultural and individual phenomenon and its explanations cannot be generalized and thus validated. This way, the mythopoeic explanation, does not transcend the bounds of the individual and cultural dimensions in its approach.

ers and algorithms that operate in the human mind. Therefore, computer models can contribute to a causal mechanistic understanding of cognitive phenomena, providing a source of insights into behaviour as well as explanations from the perspective of the functioning of underlying mechanisms (Mareschal & Thomas, 2006).

Computational tools have been producing powerful models for computing learning and development, some applied to music but mostly applied to vision, memory, face-recognition and language (Westermann et al., 2006).

The use of these models for studying the development of music information, perception and cognition, in a way that connects music and language still remains to be explored.

Accordingly, we propose to explore, throughout this dissertation, computational tools suitable for each specific research step that can best contribute to the study of factors which contribute to the shaping our cognitive structure, biasing our musical predispositions during early development. We will also explore a comparative approach to the study of early life development of musical predispositions involving both music and language, searching for possible interactions and correlations.

1.2 Dissertation outline

Music cognition is built on multiple dimensions of cognitive processing and knowledge manipulation and its study involves identifying the mental mechanisms that underlie the appreciation of music. Consequently, studying music cognition implies crossing different disciplines and establishing connections between research lines such as cognitive computing (computational developmental modelling), psychology (music cognition, developmental science), linguistics and cognitive neuroscience, leading to a trans-disciplinary analysis of the problem. This perspective brings a comprehensive and global approach to the research object that allows us to look into the problem from diverse angles of analysis.

We accordingly introduce in Chapter 2, theoretical concepts combining knowledge from different disciplines and from different perspectives, with the aim of covering pertinent theoretical issues that frame this research. The chapter is divided into four main parts. In the first part, we review key topics regarding music as a human ability, its relation with meaning, its origins, a developmental account of it and its instantiation in the brain, whether it uses specific or general domain (shared) mechanisms. In the second part, we review correlations and interactions regarding music and language, which is made necessary by the comparative approach taken by this research. In the third part, we cover cognitive processes related with this research that operate in music understanding. Finally, we identify the

study's main problem and propose the aims that determine the guideline to the developed research.

In Chapter 3, we present an overview of computational modelling, required for a research that aims to explore computational solutions suitable for studying different issues related to music cognition. Hence, in the first part of the chapter, we address cognitive computational modelling as an approach to the study of cognition. After that, we characterize computational models whose architecture is focused on neural networks, that is, connectionist models. We have applied connectionist models for modelling cognitive phenomena in two occasions in this research and thus their review becomes necessary. In the next section, we discuss development, learning and strategies for their modelling. On the following section we overview computational modelling applied to music perception and cognition. Finally, in the last section, we review the benefits and problems of computational modelling as a methodology for studying cognitive phenomena.

In Chapter 4, we examine the processes that lead to the perceptual shift of pitch representation, looking for causes that influence pitch representation throughout development. This way, we focus our research on pitch processing and representation (absolute versus relative), as well as its relations with development, from a computational perspective. Concurrently, this step is a first incursion into computational tools as a means for understanding cognitive phenomena, allowing the exploration and practice of this methodology.

With this purpose, we have developed computational simulations, based on Saffran's (Saffran & Griepentrog 2001) empirical experiments, using neural networks. More specifically, Feed-Forward Neural Networks are used in an on-line setting, and the connection weights of the network are updated applying back-propagation learning rule.

The simulations have allowed us to observe a correlation between learning and the type of pitch information being used, whereby the type of encoding influences the ability of the model to perform the task correctly and learn to discriminate between languages. These results are coherent with the findings reported by Saffran & Griepentrog (2001) in their experiments, showing a parallel between learning and the type of pitch information being used for representation. Computational simulations turned out to be a valuable tool and worthwhile to extend to further research.

Furthermore, during this research stage, we have realized that pitch representation develops depending not only on learning or age, but also on the language to which the infant is exposed (Deutsch et al., 2004). Pitch representation is affected by a learning process that involves both spoken and musical experience and thus language may be a crucial factor affecting music cognitive mechanisms. Therefore, we have hypothesized that exposure to speech and music from different cultures, with different prosodic characteristics, could result in the development of different musical representations.

Consequently, in Chapter 5, a comparative perspective has been taken, involving both music and speech. With this in view, we have conducted a study aimed at capturing rhythmic and melodic patterning in speech and singing directed to infants from two variants of the same language (European vs. Brazilian Portuguese). By taking two variants of the same language, we exclude the language itself as a variable (the lexicon) while still considering the prosodies of two different cultures.

We address this issue by exploring the acoustic features that best predict different classification problems. For the implementation of the study, we have built a data-base composed of audio recordings of infant-directed speech from two Portuguese variants and infant-directed singing from the two cultures. The computational experiments that were performed provided a detailed prosodic characterization of the infant-directed register of speech and singing from both language variants.

The findings suggest that the prosody of the surrounding stimuli, most relevant to infants (such as speech and vocal songs), is a source of rich information that provides specific cues about the rhythmic identity of their culture. Furthermore, the results point to rhythm as a key element in characterizing the prosody of language and vocal songs from a given culture. The results open up new possibilities for further extending the scope of rhythm topic, regarding infant-directed speech and singing. Can the temporal prosodic patterns specific to a given culture influence the development of rhythmic representations and predispositions? Can exposure to the rhythmic structure of speech of a specific culture cause a bias that affects music processing mechanisms and representations? Can exposure to music, or lack of it, influence language acquisition?

In order to explore these questions, in Chapter 6 we present a computational model based on rhythmic information processing and representation. Specifically, we look into the influence of exposure to the rhythmic structure of speech from a specific culture in the formation of music-related representations and consider, conversely, if the exposure to the rhythmic structure of music, or lack of it, could somehow influence language acquisition.

With this in view, we have built a connectionist model based on temporal information processing and representation and have performed computational simulations following Yoshida et al. (2010)' s human-based experiments. The model shows developmental behaviour in so far as it displays a new competence that acquired through an experience-driven transition from an early state into a later one. Therefore, the model has proven suitable for studying how the temporal prosodic patterns of a specific culture influence the development of rhythmic representations and predispositions.

Our findings reveal that exposure to the surrounding sound environment influences the development of rhythmic representations. Speech and singing directed to infants provide a means of transmitting cultural constraints, specifically rhythmic identity of a culture that is likely to condition our later

music processing mechanisms and representations.

Finally, in Chapter 7, we present the general conclusions, related to the developed research as a whole. We also summarize the contributions of this dissertation and propose directions for future research.

2

State of the art

Summary

In this chapter, we introduce theoretical concepts that were considered to be relevant in the scope of this research. The choices made to what to include here were thus intended to provide the context for the research that was developed, biased by a point of view. Addressing music cognition and its related shaping of predispositions through a computational approach requires to build a theoretical framework based on multidisciplinary knowledge. In this regard, we have structured this chapter combining knowledge from different disciplines with different perspectives.

The chapter is divided into four main parts. In the first part, we review key topics regarding music as a human capacity, its relation with meaning, its origins, a developmental account and its instantiation in the brain, whether it uses specific or general domain (shared) mechanisms.

In the second part, we review parallels and interactions regarding music and language, necessary by the comparative approach taken by this research. We explore music and language commons from a developmental and evolutionary perspective. We explore prosody as a shared characteristic between both. In addition, we review the characteristics in infant-directed speech and singing, explore common processing mechanisms and, finally, comparative studies that have been performed over music and language.

In the third part, we cover cognitive processes that operate in music understanding that have been involved in this research in any given time. These are cognitive development and categorization. The cognitive concepts are reviewed according to three perspectives: the pure cognitive domain account, the application of the cognitive process in the musical domain, and computational approaches to the cognitive process.

Finally, we identify the problem that is object of study, defining the hypothesis and research question. Furthermore, we propose the aims that guide the developed research.

2.1 Music

Music is an intriguing phenomenon of human culture, which is inherent to humans: it is a "natural outcome of a species that takes every facet of its behaviour and explores, amplifies, and extends it" (Brandt et al., 2012, p. 3). Regardless of the difficulty in defining music itself, since it means different things and has different purposes according to individual, social and cultural contexts (Nettl, 2005), we are interested in defining what musical capacity is. The human capacity for musicality is ubiquitous across human societies, present in every human being, although it can be realized in different forms, different degrees and different cultural and social environments. For Jackendoff & Lerdahl (2006), "Musical capacity constitutes the resources in the human mind/brain that make it possible for a human to acquire the ability to understand music" (Jackendoff & Lerdahl, 2006, p. 35). In this definition, it is underlying the concept of listening competence and hence music understanding. Brandt et al. (2012) proposes a complementary definition: the human ability to "engage and appreciate creative play with sound"(Brandt et al., 2012, p. 3), arising whenever sound meets human imagination. This account is flexible enough to embrace musical capacity across time and cultures.

2.1.1 Music and meaning

Music and meaning is a debatable topic that usually leads to inconclusive discussions in different fields such as musicology, philosophy, cognition and artificial intelligence. Meaning emerges from the musical experiencing and appreciation (Cross & Tolbert, 2008). However, although music is universal, its meaning is not. Across cultures, there are diverse forms in which music can be interpreted and experienced as bearing meaning. Specifically, musical meaning may arise from social domain, where there is a shared understanding (cultural grouping) and individual domain of aesthetics' foundations or even from the combination of these contexts (Cross, 2001). The relationship between music, emotions and affect, and how meaning of music carries the expression of emotions has been theorised since later seventeenth century. Music evokes strong emotional responses in their listeners, reflecting a sort of pre-linguistic symbol of emotions with social functions in communication. (Cross, 2012)

Despite the inconsistency and ambiguity in the emotional responses to music, there is a large body of theory and experiments that propose that the same type of physiological responses, as elicited emotion-producing situations can be produced when listening to music (Juslin & Sloboda, 2001). To their listeners, music can sound happy, sad, contemplative, and so forth. This association of music - which could be seen as simple sequences of tones - with particular emotions is related with acoustic cues such as mode and tempo or in the harmonic domain in terms of patterns of tension and resolu-

tion. However, these associations are historically and culturally determined (Dalla Bella et al., 2001). For example, in western music, happy is frequently conveyed with fast rhythms and major keys whether sad with slow tempos and minor keys (Hevner, 1935).

2.1.2 Music and origins

Music capacity has several open questions in debate related with its nature, origins and its evolution (Honing & Ploeger, 2012). Several disciplines have been concerned with these questions such as anthropology, ethnomusicology, developmental and comparative psychology, cognitive science, neuropsychology and neurophysiology.

There are a few hypothesis regarding its origins and adaptive significance. They address issues such as innateness and the biological determination, evolution and adaptation, and domain-specificity or general mechanisms involved in its processing (modularity). Despite of different points of view in these issues, there are converging evidences that there are universal features in music (whether as a manifestation, as system or as perception) that point to an innate constraint determination. The main generic music universals can be summarized as (1) relative pitch perception, (2) discrete pitch levels and scales existence with 5 to 7 pitches arranged within an octave range (3) octave equivalence, (4) simple frequency ratio interval notes, (5) tonality or tonal hierarchy and (6) lullabies or songs composed and sung for infants (Dowling & Harwood, 1986; McDermott & Hauser, 2005).

In respect to its nature, the musical capacity entails two related problems which are (i) whether there is an innate state of musical knowledge, prior to any musical experience within a culture and; (ii) how this initial state of musical knowledge is transformed into the adult state of music perception within a culture (Hauser & McDermott, 2003). These problems involve important issues such as innate constraints, musical experience, cognitive development and resource sharing of general mechanisms in specific domains.

Regarding the innate features, they can be explored by developmental studies that attempt to understand how infants, that lack cultural exposure, perceive music and what their musical predispositions are. However, infants never completely lack of exposure to music because in-utero experience must be considered. Humans start responding to sound around 30 weeks of gestational age and until about 6 months of age a critical period of auditory perceptual development takes place (Ilari, 2002; Graven & Browne, 2008). This shows that it is remarkably hard to control for the level of prior exposure to music. Moreover, unlike adults, infants cannot verbally report their experiences, representing an experimental challenge that leads to look for non-verbal behavioural response experimental measures. Examples of these are sucking rates for neonates when sucking a non-nutritive pacifier and looking time when oriented towards a stimulus presentation for older infants (Trehub, 2001).

2.1.3 Developmental perspective

From a developmental perspective, findings yielded developmental evidences concerning the ability of young infants to respond strongly to musical sounds (Dowling, 1999). In this sense, processing predispositions for music are assessed in infants to inspect their musical capacity (Ilari, 2002; Trehub, 2001; Dowling, 1999; Hargreaves, 1986). In their account, infants' sensitivity to musical patterns is much similar as of adults, as infants show very good auditory discrimination skills for features such as pitch, timbre, durational patterns and melodic contour. These processing skills are also used for discriminating sounds in language acquisition (Brandt et al., 2012). Additionally, infants even show more sophisticated musical processing abilities such as sensitivity to scales, harmony, musical structure and form (Trehub, 2003). However, these processing abilities gradually develop more specific features tuned to the infants' native culture. For example, at 6 months of age, infants can equally detect changes in melodies made up of pitches using both Western major/minor scale system and Javanese scales, in contrast with Western adults that more readily detect changes in melodies when using Western scales (Lynch et al., 1990). The cultural bias is also observable regarding rhythmic aspects such as musical meter and rhythmic grouping (see Chapter 6 for more) (Soley & Hannon, 2010; Yoshida et al., 2010). These evidences found early in development suggest that these abilities might be innate, a result of evolution that shaped human brains to have these specific processing skills that are required to acquire mature musical abilities (Trehub, 2000).

Regarding innate and universal preferences, most of the studies are performed in western cultures and therefore, they cannot determine if early musical predispositions are a result of the specific cultural kind of exposure and if different cultural inputs would alter preferences and perceptual discriminations. An example of this is infants' preference of consonant over dissonant intervals that, due to the impossibility to exclude probable prenatal exposure in the cultural environment, is still considered an open question (Masataka, 2006). Acculturation has crucial role and it biases music perception through the exposure to music of one's culture (Carterette & Kendall, 1999). Infants are born with a wide range of perception possibilities that, throughout development, are narrowed by means of cultural constraints (Lynch et al., 1990). This perceptual narrowing, product of the infant's cultural experience, is not specific to music, being observed as a domain-general phenomenon across perceptual modalities.

Lullabies are a musical feature that has been found across a wide range of cultures. They involve mother-infant vocal interaction and show similar forms and functionality across cultures (Trehub, 2000). Common properties in lullabies are slow tempo, and repetitive falling pitch contours (Papoušek et al., 1987). This cross cultural correspondence might suggest that infants have an innate predisposition for this musical form.

2.1.4 Modularity

In addition to innateness versus experience-dependent shaped by the environment debate, remains the discussion concerning neural specialization for music (or brain modularity) - domain-general or specific mechanisms used in music processing. These debates are specially raised concerning the origins and evolution of music. There are two main lines towards music and evolution: one that regards music as a by-product of other cognitive skills and other that considers that musical capacity is a direct target of natural selection. Whether it is a biological adaptation (Darwin, 1871; Cross, 2001; Huron, 2001) or rather a side effect of the auditory system features that evolved for other purposes (Pinker, 1997), music capacity evolved in human species and thus, it is clear that such universal capacity of music reflects changes in the brain that might have started to take place 6 million years ago (Carroll, 2003)

Music capacity may be concerning different processing modules (Peretz & Zatorre, 2005). Some processing modules might be exclusively specialized for music (Peretz & Morais, 1989), while others may be shared, for example, with speech (Patel et al., 1998). Concerning modularity, there are accounts pointing to the existence of distinct processing modules for music. Peretz et al. (2003) found patients that may go through recognition failures that affect exclusively the musical domain. This would imply the existence of specialized brain networks for musical functions, with no overlap with language or environmental sound perception networks. What still remains to determine is which are and are not the processing components uniquely involved in music (Peretz & Zatorre, 2005).

Regarding the domain-specificity, the question is to what extent music processing relies on exclusive mechanisms or whether music capacity is a product of general perceptual mechanisms that are neither music nor species specific. Musical capacity is based upon various perceptual mechanisms, some argue that are product of general perceptual mechanism and shared with most other vertebrates (Trehub & Hannon, 2006), and some argue that they are a balance between general-domain species shared and potentially unique to our species and music dedicated mechanisms (Peretz, 2006; Fitch, 2006) This discussion is especially fruitful regarding music and language domains. We will further explore this issue in the next section.

2.2 Music and Language: parallels and interactions

Music and language are two distinct sound systems with very different structural organizations that require domain-specific knowledge and representations. Language involves the manipulation of words and their syntactic properties whereas music involves representing chords and their harmonic relations (Patel, 2007). In regard to the functional level, music and language also have different communicative purposes. Regarding their functional proper-

ties, these two domains also diverge. Language is used for communication and expression of rational thought and organization of human societies, contrarily from music that lacks of semantic meaning (Meyer, 1956).

Despite their obvious differences, music and language are both ubiquitous elements in all cultures, they are human specific and both require cultural and learning transmission. Because they are both meaningful sound sequences, this resemblance invites to draw comparisons between the two domains. Music and language are both perceived primarily through the auditory system, with similar acoustic properties. They both are organized temporally, and their structures unfold over time. Regarding their structural similarities, music and language are composed by sequential elements with a specific rhythm, segmental (discrete phonemes and discrete pitches or notes) and supra-segmental information (prosody). Moreover, they are both human system constructs, based on rules and composed of basic elements such as phonemes/words and notes/chords. These elements are organized into higher-order hierarchical structures such as musical phrases and linguistic sentences, by using rules of harmony and syntax (Besson & Schön, 2001).

Music and language similarities have intrigued a wide range of thinkers, including musicologists, linguists, biologists and philosophers. Their correspondences have been explored since their origins, with Darwin's (Darwin, 1871) hypothesis indicating a shared evolution history. Throughout the next sub-sections, we will explore music and language's connections from different points of view.

2.2.1 Common origins

The possible connections between music and language's evolutionary pathways are a topic of debate involving anthropologists, psychologists and scientist of animal behaviours. There is a hypothesis that points to a shared evolutionary history of these two domains. During the evolution to bipedalism, caregivers had to somehow compensate the lack of physical contact when infant-riding was lost. This compensation was made through the development of special vocalizations that also served to better co-ordinate mother-infant interactions. This interaction would eventually provide the infant the opportunity to acquire the capacity to learn vocal usage (Falk, 2004). These vocalizations that might have evolved from this need of compensation represent a communication system or "prosodic protolanguage" that provided a precursor for music and language as we know them today (Masataka, 2009).

Moreover, these vocalizations might have been the origins of infant-directed speech and singing, that are used across all cultures (see Section 2.2.4). Speech and singing directed to infants are still used by its affective salience and attention-getting properties to communicate emotionally with infants on a basic level. In turn, the infant-directed register is preferred by infants relative to adult-directed register (Fernald, 1985; Cooper & As-

lin, 1990; Masataka, 1999). This special parenting style is also known to work, at the prosodic level, as a beneficial facilitator of language learning for preverbal infants (Thiessen et al., 2005).

2.2.2 Developmental perspective

From the infant perspective, in a preverbal stage, speech and singing might not be as differentiable as for adults and young children given that, without linguistic knowledge, both domains might be perceived as sequences of sounds that unfold in time. Thus, at an early stage of development, infants might be attending to melodic and rhythmic properties of speech in the same way as in music, sharing resources for processing both (McMullen & Saffran, 2004).

Brandt et al. (2012) go further and describe spoken language as a special type of music. In their view, spoken language is presented to the child as a vocal performance from which infants extract musical features first. Without the ability to hear musically it wouldn't be possible to learn to speak.

Hence, at an early stage, musical and linguistic abilities may be based upon the same learning competences. This way, the musical aspects present in speech and singing directed to infants become an integrant ontogenetic factor in the development of human communication capacities developing, at an early stage, an intermediate communication ability based on prosody.

2.2.3 Prosody

Prosody is a central common supra-segmental cue in music and language. The term prosody has its origins in ancient Greek culture, where it was originally related to musical prosody (Nooteboom, 1997). Musical prosody can be seen as the musician's manipulations of acoustic signals to create expression, communicate emotion and clarify structure. Thus, in order to shape music, the performer adds variations to the sound properties, including pitch, time, amplitude and timbre (Palmer & Hutchins, 2006). The manners in which performers' model musical pieces in order to add expression is very similar to the ways in which talkers manipulate speech sounds and sequences (i.e., lengthening or shortening sounds, adding modulations, making attacks more or less abrupt, etc.). This way, both musical and speech prosody relate to the manipulation of acoustic features to convey emotional expression and to provide segmentation and prominence cues to the listener.

Speech prosody refers to speech properties that go beyond sequences of phonemes, syllables or words, that is, the supra-segmental properties of speech. These characteristics comprise controlled modulation of the voice pitch, stretching and shortening of segments and syllable durations, and intentional loudness fluctuations (Nooteboom, 1997). In other words, prosodic cues are associated with acoustic variations of pitch or fundamental frequency, spectral information or voice quality, amplitude, and relative durations in patterns of speech.

In respect to the common prosodic structure, there are comparable rhythmic and melodic features in music and language worth to be analysed. As for rhythmic features, both music and language are composed by systematic temporal accentual and phrasal patterns (Patel, 2008). However, music rhythm is usually regular, with isochronous pulse and perceptually periodic (Bispham, 2006; Fraisse, 1984), whereas in speech there are no repeating patterns temporally regular (Port, 2003). Comparing music and language as bridging non-periodic aspects of both has shown fruitful (Iversen et al., 2008; Yoshida et al., 2010) and, thus, periodicity should not limit the study of rhythmic correlations between both.

Regarding melodic aspects, both music and language can be seen as "an organized sequence of pitches that conveys a rich variety of information to the listener" (Patel, 2008, p. 182). In speech, the melody carries affective, syntactic, emphatic (signalling prosodic grouping) and emotional information. "From a musical perspective, speech is a concert of phonemes and syllables, melodically inflected by prosody" (Brandt et al., 2012, p. 4). Musical melody also conveys information, evoking a rich set of perceptual relations. The human perceptual system converts the sequence of pitches into perceived relations such as interval structure, grouping structure and tonality relations (hierarchical pitch relations and harmony) (Patel, 2008). These perceptual associations may lead to many more meta-relations, resulting in a psychologically distinct perception of musical melodies in terms of subjective experience.

2.2.4 Infant-directed speech and singing

Prosodic features in both music and language are highly salient to infants (McMullen & Saffran, 2004). The sensitivity to affective prosody is present in infants since the first days of life (Cheng et al., 2012). This early tuning for prosodic information might start in utero experience, where patterns of rhythm, stress, intonation, phrasing and contour are the available human-produced external sound source. At the same time, the prosodic stimulation that comes from the infants' environment is highly rich from musical and linguistic domains.

Prelinguistic infants, since they are born, are exposed to a special register of music and speech, known as infant-directed singing and speech. In these registers, both domains are subjected to adjustments by caregivers in order to be the most attractive to infants. Infant-directed speech' modifications involve the use of exaggerated intonation contours, expressed in slower rate of speech, higher fundamental frequencies, wider range of pitch variations (Papoušek et al., 1987) and characteristic repetitive intonation contours (Papoušek & Papoušek, 1991; Papoušek et al., 1990). These modifications are found cross-culturally (Fernald, 1992; Falk, 2004) and are used to regulate infants' emotional state and convey meaning, e.g., approval and disapproval. The emotional content of infant-directed speech can be even understood

across different cultures and languages (Bryant & Barrett, 2007). Additionally to the functional role in emotional communication, the modifications in infant-directed speech might aid infants' language learning (Thiessen et al., 2005).

In respect to infant-directed singing, caregivers also modify their register when they interact musically with infants. The repertoire is composed by a few songs, play-songs and lullabies, normally characterized by simple and repeated pitch contours and these properties are also cross-cultural (Trainor et al., 1997).

2.2.5 Modularity versus shared resources

As a result of the early developmental trajectory of musical and linguistic experiencing, it is still not clear how these two domains are instantiated in the brain, whether if similar processing mechanisms such as memory and learning are shared in knowledge acquisition in the two domains. This brings back the discussion about the modularity of mind and to what extent are cognitive mechanisms specifically dedicated to particular domains but, this time, specifically focusing on music and language domains. In Fodor (1983)'s point of view, music and language processing makes use of distinct architectural brain regions that are independent in such way that there is no connection between them. It has been observed that the perception of speech and music elicits different patterns of activation in the brain. Speech perception relies more on the left hemisphere and music relies more on the right (Callan et al., 2006). This lateralized activity has been related with the use of different types of processing, with speech demanding for very rapid temporal processing required for the extraction of segmental and phonemic information and music associated with pitch features that vary over longer time windows (Zatorre et al., 2002).

On the other side, recent scientific research has been exploring music and language possible connections in a comparative context, yielding cues in the sense of neurophysiologic, perceptual and cognitive overlap. Recent studies have been suggesting significant overlap in neural regions underlying music and speech perception (Merrill et al., 2012). The overlap may respect to mechanisms such as acquisition (Schön et al., 2008), encoding basic auditory cues (Kraus & Chandrasekaran, 2010), detecting violations in predicted structures (Slevc et al., 2009) and implicit memory (Ettlinger et al., 2011). Moreover, there are evidences showing that the activation to infant-directed speech and to instrumental music show significant overlapping in newborns (Kotilahti et al., 2010). These findings suggest that the hemispherical separability observed in adults might emerge over the course of development. Finally, the same temporal precision has shown to be necessary to process both music and language (Hukin & Darwin, 1995), with small time windows being crucial for timbre recognition in music (Hall, 1991) and speech processing relying on longer time-scale windows that correspond

to syllable-sized vocalizations (Morillon et al., 2010). The observed hemispherical differences might be a result of a cortical specialization in aspects of general auditory processing rather than specialization for music or speech (Brandt et al., 2012).

In order to reconcile contradictory evidences in respect to music and language correspondences, Patel (2007) proposes a theoretical framework that aims to make a distinction between domain-specific knowledge and shared neural resources. The resource-sharing framework is based on two main principles that involve domain-specificity and neural overlap concepts: (i) music and language entail domain-specific representations and; (ii) the brain shares neural resources between music and language processing, which operate similar cognitive processing upon domain-specific representations. In other words, the processing algorithms would be shared but the knowledge outcomes, memories and sources would be separated.

2.2.6 Comparative studies

Music and language cognition and its interactions have been addressed with diverse scientific approaches. The contour-based processing in early infancy experience may have an effect on prosodic predispositions and sensitivity. The notion that linguistic environment influences musical culture whether in terms of rhythm or in pitch representation has motivated empirical comparative studies that explore these possible correspondences.

Deutsch (1991) explores the influence that language can have on music perception by testing the differences in the perception of musical patterns known as the tritone paradox. The conducted study is led by the hypothesis that the perception of the tritone paradox might be related to the processing of speech sounds. The study demonstrates that spoken language is the basis factor that influences the individual differences in perception. Krishnan et al. (2005) also found evidences on language's influence on pitch neural encoding. In their study, findings showed that tone language speaking subjects (Mandarin Chinese) had stronger pitch representations and smother pitch tracking than the English speaking subjects. These results imply a relationship between pitch representation and the long-term experience with language, revealing an experience-dependent auditory neural plasticity of mother tongue. In later research, pitch representation is revisited, showing that although absolute pitch representation might be present at birth universally, tone language infant learners may maintain this kind of representation due to their language learning requirement to associate pitches with verbal labels during the critical period for speech acquisition, whereas learners of other languages miss that capability (unless it is maintained by some musical training) (Deutsch et al., 2004). Pitch representation is further elaborated in Chapter 4.

Regarding rhythm, there are also evidences that support music and language relations. Patel et al. (2006) conducted a study where they compared,

using quantitative methods, musical and spoken rhythm from British English and French. Their findings support the notion that the prosody of a culture's spoken language is reflected in the rhythmic and melodic structure of its instrumental music. These results are reinforced by Hannon (2009) who, in his experiments, found that listeners perceive language-specific rhythms in musical contexts and classify instrumental sequences based on the language rhythmic information. Other example comes from Soley & Hannon (2010) whose experiments suggest that infants' meter preferences might be driven by culture-specific experience.

Moreover, rhythmic grouping, that was viewed as governed by universal perception principles, was found to be dependent on auditory experience (Iversen et al., 2008). Specifically, the language-dependent bias in perceptual grouping is developed by 8-months when linguistic phrasal grouping develops, and the developed preferences are consistent with the mother tongue's structure of infant subjects (Yoshida et al., 2010). The subject of language influence on rhythmic representations is further developed in Chapter 6.

The studies that were presented represent a body of research that is still underdeveloped. However, they reveal cognitive interactions and open a promising path for comparative studies that aim to explore music and language relations (Patel, 2008). In this line, we consider that the comparative approach that we take, involving music and language cognitive processing, allows the research to take advantage of the body of knowledge produced in the language processing, either in cognitive research or in computational modelling techniques. The crossing of information that comes from two fields introduces a vaster character to the research, enabling to correlate music and language systems, and opens new possibilities for building more solid hypothesis related to its correlations and interactions. Moreover, using speech and vocal songs material throughout the research might add qualitative extent to the research object and brings validity to the research.

2.3 Cognitive processes that operate in early music understanding

2.3.1 Cognitive development

The individual development that can take place within a life span of an organism is an active process of change by which proper biological structures and abilities emerge in each individual by means of complex, constructive and variable interactions between endogenous (genes) and environmental factors (Johnson & Hann, 2011). The nature of the interaction between genes and environment remains controversial. The paradigmatic nature-nurture debate describes two dimensions, the Nativist versus the Empiricist views on cognitive development. Radical Nativists defend the idea that almost all knowledge is available to the infant before any experience. Radical

Empiricists, in contrast, defend all knowledge is acquired through experience with the learning abilities which infants are born with.

The relations between learning and development and the nature of their interaction is also an issue of debate (Lindner & Hohenbergen, 2009). There are three main lines in this debate that can be defined: (i) the interaction between development and learning is unidirectional, where learning capitalizes on the achievements of development and cannot happen unless a certain stage of development has been accomplished (Piaget, 1953); (ii) the interaction is bidirectional wherein learning and development mutually influence each other, by development being able to enable or limit learning and, in turn, development progresses with learning (Kuhl, 2000); (iii) There are no boundaries between learning and development, being both part of a dynamic system in a continuum process of change (Thelen & Smith, 1994).

Learning is experience-dependent and results from practice and exercise that produces a relatively permanent change in organisms (Gordon, 1989). Learning can occur in implicit fashion, where the acquisition of knowledge takes place independently of conscious will to learn and without the presence of explicit knowledge about what was acquired (Reber, 1989). Explicit learning, in contrast, differs in the aspect of awareness, where learning is triggered by the presence consciously accessible knowledge. Hence, implicit learning works as the default mode mechanism (Reber, 1989).

Statistical learning can be considered a form of implicit learning, characterized by a discovery procedure that detects basic statistical properties such as patterns and regularities and results in the acquisition of structure in the environment. This type of learning mechanism is neither species nor domain specific and thus considered to be basic and robust. It is commonly related to language learning in young children, regarding the discovery of prosodic regularities and word segmentation (see Chapter 4 and Chapter 6 for an application of this learning mechanism) (Saffran et al., 1996; Saffran & Griepentrog, 2001; Jusczyk et al., 1994).

There has been a traditional dichotomization regarding the progression of cognitive development. In one side, there are the ones defending that development occurs in discontinuous stages that are universal commons and characterize all domains of development (Piaget, 1971). This definition focuses on the structure or the sequence of development that is necessary for going from initial to adult form. A branch of study derived from this approach to development is the demarcation of transitions between stages, specifying how progressions are stage-like.

One criterion taken, beyond age, was the qualitative changes that mark transitions to new stages, wherein the structural organization is analysed, rather than the amount of a behaviour or capacity. However, with this criterion, every increment of learning would mark a new stage (Fischer et al., 1984). For this reason, and to bypass this problem, researchers used an intuitive sense for determining what an important qualitative change is. Con-

servation[1] is one example of a structural definition that has been provided to specify important changes (Biggs & Collis, 1982).

On the other side, cognitive development is a continuous process, showing different individual developmental patterns driven by particular experiences or environmental effects (Feldman, 1980). This perspective is characterized by a mechanistic approach to development, emphasising the functions that serve behavioural change. This way, individual differences in development are assessed concerning that environment influences behaviour through the principles of learning or problem solving and thus, development is as variable as its context (Skinner, 1969). Consequently, developmental paths that are specific to domains and experiences, are potentially infinite in their diversity.

However, this dichotomization has been refuted by data that demonstrate that both positions coexist, showing that development occurs in stages and also shows great individual diversity (Fischer & Silvern, 1985). There are evidences showing that cognitive development has stage-like changes characteristics with cross domain consistency and, at the same time, is characterized by environmental effects and individual differences. This way, the nature of cognitive development emerges from a combination of environmental and organismic factors. This integrated perspective determines that development is plastic, varying in response to environmental variation but, at the same time, this diversity is constrained by the organization of developmental stages (Gollin, 1984; Lerner, 1984). Thus, given different contexts, the same initial structural conditions will produce different probable developmental paths and outcomes, creating a probabilistic epigenesis. Therefore, individuals exhibit developmental plasticity, functioning differently in different contexts. These contexts may be social group and cultural organization. In turn, the functioning influences the subsequent structures that also influence the subsequent functioning factors, such as genetics, behaviour or neuroanatomy. The interaction between environmental constraints and cognitive structure will be developed in Chapter 6.

Brain goes through a maturation process that involves different types of neural change, to acquire specific cognitive acquisitions. During an initial stage of development, especially between the late prenatal and early postnatal period, infants' brains go through a synaptic overproduction that tends to gradually decrease to adult levels after the age of two. This high density of production of synaptic connections shapes the prefrontal cortex and other related brain areas, and it is directly involved in forming stable representations capable of being accessed and used on-line (Goldman-Rakic, 1987).

The heighten levels of brain plasticity occurring in simultaneous with critical periods exert a great influence in shaping the neural circuits and form enduring representations, through experience, altering permanently perform-

[1] Conservation refers to a psychological task used to test a child's ability to capture the invariance of an object property (such as substance, weight or number) after it undergoes physical transformation.

ance and behaviour. Critical periods are developmental time windows during which a specific type of experience or its deprivation has a prominent effect on the development of an ability or behaviour. Moreover, the experience or its absence during a critical period can cause the consolidation of structural modifications in the architecture of neural circuits, leading to the stabilization of certain patterns of connectivity that become energetically preferred (Knudsen, 2004).

The development of musical expertise might be dependent on critical periods (Trainor, 2005). The analysis and definition of musical critical periods still has no simple answer due to the complex nature of musical structure and the blurry identification of general and musical-system-specific learning mechanisms involved in perception and cognition. However, there are some identified examples where early experience might have a permanent effect, such as the development of the auditory cortex and the tonotopic map formation (Moore & Guan, 2001; Weinberger, 2004) and the development of pitch representation (absolute versus relative) (this issue is further analysed in Chapter 4). Trainor (2005) considers that it is of essential importance the analysis of the mechanisms that underlie the interaction between genetic and experiential factors that create musical critical periods. This analysis would contribute to a better understanding of the musical critical periods as well as their timing and duration.

Despite how cognitive development progresses, it is clear that development is to change. Thus, for a comprehensive understanding of development, it is necessary to first understand the mechanisms that produce that change, contributing for cognitive development (Siegler, 1989). By cognitive-developmental mechanisms it is intended to be interpreted as any mental process that improves the ability of children to process information. In the regard of looking to understand how developmental mechanisms operate, some of the best ideas were achieved in the context of connectionist models, for example, regarding associative competition. The connectionist approach has consequently influenced the thinking about cognition. In Chapter 3 we further explore connectionist models.

2.3.2 Categorization

Categorization is a mental process that requires some form of abstraction or generalization, reducing the complexity of a continuous world of sensory features, by partitioning them into equivalent classes. In other words, it is the ability to relate familiar experiences to each other and to other novel experiences by focusing on common aspects of information and ignoring differentiable features. The recognition of the same object under different circumstances involves learning a property of an item that is extended to other similar items (Sloutsky, 2010). This ability is considered essential for knowledge acquisition. Besides, the capacity to build coherent mental representations or category representations for similar or like entities

provides the organization and stability of cognition (Quinn & Eimas, 2000) and reduces the load on memory (Rosch, 1975).

The way items are grouped or categorized determines how the relationships between the objects are learned and how these relationships are generalized to novel items (Mareschal & Quinn, 2001). Two types of object categorization have been identified in infancy: perceptual and conceptual categorization (Mandler, 2000). Perceptual categorization is considered to be an automatic part of perceptual processing that computes perceptual similarity between objects, creating perceptual representations of the object. In contrast, conceptual categorization processes focus on what objects do, forming category representations based on objects' functionality (Mandler, 2000). In general, perceptual categorization is applied in object identification and conceptual categorization in inductive inference, although there may be interactions between the types of categorization. For the scope of this dissertation we will only refer to the perceptual categorization type.

Early categorization has been studied seeking to understand how the category representations emerge during development, how they are formed and how they develop (Younger & Gotlieb, 1988). Infants display the ability to form perceptual category representations since their first days of life (Slater, 1995). They show flexible and responsive processes in category formation that adapt to the variability characteristics of the stimuli (Bomba & Siqueland, 1983).

The mental representations of categories may form hierarchical organized systems with different levels of inclusiveness. Global categories may be formed earlier than basic level categories due to the great efficiency of their representations and the more discriminable and frequent attributes that characterize the global categories (Quinn, 2007). With the increasing frequency of experience, items will tend to be represented in more differentiated and subordinate levels, forming more basic categories.

In the musical domain, categorization is critical for music processing and it is involved in several tasks. For example, in early development, infants possess specific representations relative to tempo and timbre, not being able to generalize music with changes in these attributes and recognize pieces played at different tempo rates or new instruments (Trainor et al., 2004). As a function of experience, the level of specificity changes and the ability to represent music abstractly grows. Another example is the organization of rhythmic patterns in music which relies in the categorization of durations according to the hierarchical temporal structure of the music. This metrical categorization process is culturally biased by typical duration ratios, since infants show flexible perception of meter but, in contrast, adults are tuned to the metrical categories of their musical culture (Hannon & Trehub, 2005). In chapter 6, we address the formation of category representations relative to temporal prosodic patterns that are present in infant directed speech and songs from a specific culture.

Categorization has been addressed by the computational perspective

(Kruschke, 2008). Computational models have been used to explore the mechanisms that underlie the category learning in infancy (Marechal & French, 2000). Marechal & French (2000) conclude that both infants and the connectionist model used covariation information to segregate items into different categories. Moreover, they suggest that categorization emerges from the interaction between the mechanisms internal to the subject (infant) and the properties of the environment (the stimuli).

Throughout this dissertation, in the different research works performed, we have relied on categorization in our computational approach. Especially in chapter 5, where different categorization problems are posed in order to capture characteristic rhythmic and melodic patterning in infant directed speech and songs from two different cultures that share the same language (Portuguese).

2.4 Problem statement

The study of how the brain processes music emerged as a rich and stimulating area of research in cognition, perception and memory. Experimental psychology and neuroscience show results pointing that the development of musical representations and predispositions in infancy, whether concerning pitch or rhythm features, depend both on experience and language (Deutsch, 1991; Saffran & Griepentrog, 2001; Krishnan et al., 2005; Iversen et al., 2008; Hannon, 2009; Soley & Hannon, 2010; Yoshida et al., 2010).

However, there are still many open issues regarding the processes and mechanisms that underlie music cognition. The above referred disciplines identify and describe the cognitive phenomena. Yet, their account regards on a perspective based on the features of human behaviour which, in turn, are an emergent result of the inner working processes and mechanisms that operate in the human mind. The understanding of human musical information processing is a very fertile area, with growing research interest and that still has a lot of potential for further developments. Additionally, the ubiquity and, at the same time, diversity attributes make musical capacity a rich field. At the same time, music processing engages a series of complex perceptual, cognitive and emotional mechanisms and has an important role in ontogenetic development and human evolution. All these attributes make music an ideal means to study human cognition and can contribute with new insights for the understanding of the human brain (Koelsch, 2012; Pearce & Rohrmeier, 2012)

In parallel, computational modelling has produced powerful tools for computing learning and development. There are models that hold features that embody basic characteristics of the human brain such as self-organization, plasticity or experience-dependent structural elaboration. These models have proven to be a powerful tool for problem solving in pattern recognition, prediction, and associative memory (Haykin, 2009). They have also demonstrated success as being suitable for solving cognitive developmental modelling problems (Elman, 2005; Munakata & McClelland, 2003; Quinn & Johnson, 1997; Westermann et al., 2006). Most applications of these models are in vision, memory, face recognition and language (Westermann et al., 2006). However, the use of these models for studying the development of music information, perception and cognition, in a way that connects music and language still remains to be explored.

At this point, we hypothesize that the infants' development of musical predispositions is influenced by the prosodic features that are present in the sonic environment of their culture. Consequently, this hypothesis, poses the following question: how do these features or elements influence the development of musical representations and predispositions during infancy?

Regarding this question, the purpose of this research is to explore computational solutions suitable for each specific research stage (i.e. Chapter

4, Chapter 5 and Chapter 6), that can best contribute for the study of the shaping of the human cognitive structure, biasing the musical predispositions during early development. This research will also explore a comparative approach to the study of early development of musical predispositions that involves both music and language, searching for possible interactions and parallels.

2.5 Aims of the study

The capacity to endow meaning to music is found in every human being, in every culture (Nettl, 2000; Cross, 2012). From the musical experience, meaning emerges differently in each human being (Cross, 2001). However, the interpretation of musical meaning can have common factors within cultures, showing that experience might bias the way humans understand music (Deutsch, 1991; Krishnan et al., 2005; Iversen et al., 2008; Hannon, 2009; Soley & Hannon, 2010; Yoshida et al., 2010). During infancy, the brain experiences a high level of brain plasticity period, which is critical in shaping the neural circuits and form enduring representations and altering permanently performance and behaviour. During this period, developmental processes are especially sensitive to environmental input, and its experiencing (or the lack of it) has strong influence on the consolidation of structural modifications in the architecture neural circuits and thus on the acquisition of adult level abilities in specific areas (Knudsen, 2004). Among the auditory information to which infants are exposed, the most salient are speech and singing sounds (Masataka, 1999; Werker & McLeod, 1989). From the perspective of a preverbal infant, music and speech may be not as differentiated as they are for older children and adults (Brandt et al., 2012). They may be perceived as sound sequences that unfold in time, following patterns of rhythm, stress and melodic contours (Masataka, 2009).

In this context, the main goal of this research is to explore, relying on computational modelling techniques, factors that contribute to shape our cognitive structure, influencing our predispositions and representations that allow us to enjoy music and make sense of it as it is heard. This goal takes in the operational objectives described following:

- Investigate the factors that lead to the perceptual shift in pitch representation.
- Explore and identify the prosodic features that best characterize and provide specific cues about the cultural identity of the auditory environment of an infant by comparing rhythmic and melodic patterning in infant directed speech and vocal songs of two cultures.
- Study how the temporal prosodic patterns of a specific culture influence the development of rhythmic representations and predispositions.

- Apply a comparative approach that involves both music and language cognitive processing in order to explore its interactions and possible parallels.
- Build and test computational models for pitch and temporal processing by running experiments, using empirical data from experimental psychology.
- Produce, based on the architecture, structure and data derived from the models, explanations for how the sonic environment of a specific culture influences the development of our musical representations and predispositions.
- Contribute to the building of a theoretical framework, based on multidisciplinary knowledge that allows a comprehensive approach to the elements that influence music cognition and perception.

3

Methods

Summary

In this research, we propose to explore computational tools suitable for studying different issues, yet all commonly related with the central problem: exploring factors that contribute to shape the human musical representations and predispositions. Hence, we devote this chapter to the methods followed, that is, computational modelling. The chapter is divided into two main parts: An overview of computational modelling and a summary of the methodology followed with computational tools. In the first part, we address cognitive computational modelling as an approach to the study of cognition. After that, we characterize computational models whose architecture is focused on neural networks, that is, connectionist models. We applied connectionist models for modelling cognitive phenomena in two occasions in this research and thus its review provides an introduction to the unfamiliar reader. In the next section, we discuss development, learning and strategies for its modelling. In the following section we overview computational modelling applied to music perception and cognition. Finally, in the last section, we review the benefits and problems in computational modelling as a methodology for studying cognitive phenomena.

3.1 Computational modelling overview

3.1.1 Computational cognitive modelling

Computational models are algorithms that can be implemented as computer programs. These programs can run, be manipulated and tested. Thereby, computational models can be useful theory-building tools, as they can incorporate the description of cognition in computer algorithms and programs (Turing, 1950). In this perspective, they can be taken as theory representations. In this sense, computer models as theories can be divided into product-based theories or input-output theories and process-based theories (Sun, 2008).

Product-based theories focus on the result of a process but do not commit to a particular mechanism or process. Therefore, these theories do not make any predictions about the process that is involved in producing the result. This way, the evaluation of these theories can be performed by simply measuring the result of the process.

The contrasting process-based theories aim to understand and specify, in an accurate way, the computational models' representations, mechanisms, and processes. The models, thus, explain how human performance occurs and by what mechanisms, processes and knowledge structures. Thereby, computational cognitive modelling explores the essence of cognition and various cognitive functionalities through the analysis of computational models of representations, mechanisms and processes. However, evaluating a process-based theory is not simple, because it involves using process measures, if they are available and valuable, valid and relevant. In this regard, computational models hold a great advantage, since they allow inspecting their internal representations.

Regarding cognitive science, there may be 3 categories of models, namely, mathematical, verbal-conceptual, and computational models (Bechtel & Graham, 1998). Mathematical models are about relationships between variables, using mathematical equations. These models can be viewed as a subset of computational models because they could hypothetically lead to computational implementations, but they are commonly sketchy and lack process details. Verbal-conceptual models describe entities, relations, and processes in informal natural languages. But language often fails in the attempt to capture complexity, richness and subtlety of phenomena. Computational models, in turn, present process details using algorithmic descriptions. The explicitly to which the computational perspective forces often leads to new ways of understanding observed phenomena and a complementary view to empirical experimental methodologies in understanding cognitive phenomena.

The computational approach to cognition is constantly challenged by the extreme complexity of the systems to be analysed. For this reason, the view of hierarchical levels of analysis was introduced, to deal with complexity

(Marr, 1982; Newell & Simon, 1976; Sun et al., 2005). This implies that computational models can be used to address different levels of analysis, that is, levels of abstraction. Following this notion of levels of analysis, Marr (1982) proposed three different levels at which an information processing task carried out by any device must be understood. These levels are: (i) the abstract computational theory of the device, where what the program does is specified, without worrying about exactly how the program does it; (ii) the choice of a representation and algorithm that manipulates and transforms the representation in order to implement the computational theory and; (iii) the implementation level, or how the algorithm and representation are realized physically. Despite the independence that exists among these levels of description on understanding information processing, they remain connected because the options taken in one level can condition at a certain degree the options taken on the other two and, in this sense, they are logically and causally related.

The risk that Marr's approach might take is that the implementation level could be irrelevant faced to the other two levels of analysis as if the implementation issues wouldn't affect the algorithmic and computational theory levels. Indeed, computer algorithms can become implementations by the automatic process of compilation. However, the brain functioning and its neural implementation are a complex matter to express at a higher-level of description and, thus, it is not obvious that they do derive automatically from a higher-level. This automatic derivation is not obvious too in parallel computing, where new and unexpected behaviours can emerge from the implementation of the higher-level of computational theory. Given the complex nature of the brain, its parallel information processing functioning and, consequently, its emergent cognition, it is dangerous to aim simple explanations that are framed in the operating mode of standard computers.

On the other hand, an approach that emphasises the implementation level, reducing the importance of the computational theory and algorithmic levels is an approach that might lead to a poor understanding of the properties of the phenomena to model and the degree of relevance of these properties. The poor specification of the goals, purposes and constraints of a cognitive process can lead to complicated models that offer little account on the cognitive phenomena the model aims to explain.

Consequently, a balanced approach between all levels of analysis, that creates links between information across all levels can be much more valuable. An approach that stands on the trade-off of between a simple model and the aspiration to include as much of the mechanisms that are known from the cognitive phenomena. This is where we situate the approach that we have pursued along this research.

3.1.2 Connectionism

In this section we characterize computational models whose architecture is focused on neural networks. These models are known as artificial neural networks (ANN), connectionist models or parallel distributed processing (PDP). Connectionist models were stimulated by research produced on how the brain processes information. They can gain considerable diverse forms or architectures, but all models are essentially constituted by the same basic components: simple processing units, linked by weighted connections. In the networks, processing is distributed, characterized by patterns of activations across the processing units. For this reason, this approach for studying cognitive phenomena is called connectionism. This approach has been establishing as a valuable and contributing tool for the study of cognition over the last twenty years (Thomas & McClelland, 2008). Connectionist models have been applied to several cognitive skills (Houghton, 2005) such as memory, attention, perception, language, concept formation and reasoning. Many of these models concern the adult cognitive performance. However, these models provided a tangible means to observe their internal representations and how they evolve over time. The consequence of this was an increasing focus on developmental phenomena and the origins of knowledge (se section 3.1.3). For now, we will present a short description of this category of models.

Basic concepts

Parallel distributed processing models emerged from the pursuit to break with the formal manipulation of symbolic expressions paradigm, inspired by biological neural networks functioning (Hinton, 1989). Seeking to understand human cognitive capabilities led to an understanding of how computation is organized in systems like the brain, that consist of substantial numbers of slow processing nodes that are interconnected. This way, connectionist models are composed by simple neuron-like processing units that interact trough weighted connections. Each unit has a state that is established by the input received from other units in the network. The main goal of connectionism is to contribute with efficient learning procedures that permit the models to build complex internal representations of their environment. Connectionist models have numerous variations and architectures. Rumelhart et al. (1986) propose a general framework enabling the discussion of various connectionist models by identifying eight major aspects of parallel distributed processing models. By its pertinence and utility in understanding these models, we will describe them very shortly next.

- **A set of processing units u_i.** These units can be distinguished into input, output and hidden units. Distributed representation means, thus, that one in which the units represent a small feature-like entity. The meaningful level of analysis becomes then the pattern as a whole of the units.

- **A state of activation** $a_i(t)$**.** At each point in time t, each unit has an activation value. The vector $a(t)$, composed of N real numbers, represents the pattern of activations over the set of processing units, that captures what the system is representing at time t. It is possible, this way, to register the system's processing as the evolution of a pattern of activity of the whole units, throughout time.

- **An output function for each unit** such as $f_i(a_i(t))$ that maps the current state of activation $a_i(t)$ to an output signal $o_i(t)$. This is how units interact, transmit signals to their neighbours. Thus, their degree of activation determines the degree to which they affect their neighbours. Commonly, f is a threshold function and thus a unit does not affect its neighbours unless its activation surpasses a certain value. It is also common that the output of the unit depends probabilistically on its activation values and f is assumed to be a stochastic function.

- **A pattern of connectivity among units.** The absolute value of the weight w_{ij} represents the strength of the connection between unit u_i and u_j. Positive numbers for w_{ij} indicate an excitatory connection between u_i and u_j and negative values of w_{ij} signify an inhibitory connection between u_i and u_j. The matrix W, that contains the weight values for the network units, represents the pattern of connectivity in which the weight w_{ij} stands for the strength of the connection between u_i and u_j. The pattern of connectivity contains the information that determines the response yielded to any given input. This way, this matrix is the structure holding the "knowledge" gained by the system with respect to a given task or problem.

- **A propagation rule** for propagating patterns of activities throughout the network. This rule combines the output values of the units, represented in vector $o(t)$ and the values of the connectivity matrix W in order to produce a network input into each receiving unit.

$$net_i = W \times (t) = \sum w_{ij} o_j$$

- **An activation rule.** This rule determines how the inputs that flow into a given unit are combined within each other and the current state of that unit to produce its new level of activation. F is the function that derives the new activation state by taking $a(t)$ and the vectors net_i for each type of connection

$$a_i(t+1) = F(net_i(t))$$

- **A learning rule** whereby patterns of connectivity are modified as a function of experience. Changing the knowledge structure in connectionist models entails the modifying of the interconnectivity patterns.

In this sense, these systems gain plasticity given that their patterns of interconnectivity are always changing by weights modifications as a function of experience. These modifications can be of three types: i) the construction of new connections (constructivism); ii) the deletion of existing connection and; iii) the modification of the strength of the existing connections. Learning rules can have numerous variations but fundamentally they descend from the Hebbian learning rule (Hebb, 1949). The basic idea behind it is that the strength of a connection strength is a result of the synaptic neural activity. This means that the weight between two units should be changed proportionally to the activity between those two units. This idea is expressed by

$$\Delta w_{ij} = \eta a_i a_j$$

where η is the constant of proportionality or also called learning rate. This is mostly valid for rules that concentrate on the modification of connection strengths (i.e. the case iii)).

- **An environment** within which the system must operate. In these models, the environment is represented as a time-varying stochastic function over the space of input patterns.

Neural Plausibility

Neural networks were built on the parallelism with the computational properties of neural systems and roughly capture some principles operating on biological neurons. Similar to the human brain, that contains near 10 billion neurons in which each one contributes its part to overall human cognition, neural networks or parallel distributed processing systems are composed by simple processing units that contribute for the overall process. This way, parallel distributed systems can be a valuable tool for understanding how collective interactions between several processing units such as neurons, can lead to the emergence of cognition (O'Reilly & Munataka, 2000; O'Reilly, 1998).

Neural networks are flexible systems that embody the plasticity properties of human brain, integrating the capacity to learn and adapt through experience. Connectionist models allow conforming to brain-style computation, providing a neurological plausibility that is absent from other modelling methods (Shultz, 2003). Moreover, computational models that are based on biological properties of the brain can contribute for understanding all of its complexity.

However, this view is not unanimous and neural plausibility of neural networks is put in question (see Thomas & McClelland (2008)). Thomas & McClelland (2008) claim that "neural plausibility should not be the primary focus for a consideration of connectionism. The advantage of connectionism,

according to its components, is that it provides better theories of cognition" (Thomas & McClelland, 2008, p. 29).

Connectionist contributes for cognitive science

Connectionism represents an opportunity to centre the attentions on causal explanations of cognitive mechanisms, focusing on how cognitive capabilities work, rather than a descriptive classification of what are the cognitive capabilities and when they appear through development. In connectionist models, it is possible to explain the nature of the representations that govern the behaviour of the network. This happens because these models offer an account of the representations that trigger the performance on a cognitive task, allowing looking into the model's representations and providing a means of explanation for the mechanisms that underlie the model's behaviours. Exploring the nature of these underlying representations is fundamental for contributing on the "how" perspective of understanding in these systems, focusing on an explicative rather then descriptive perspective on cognitive phenomena. This is due to, as Hawkins & Blakeslee (2005) states, the input-output paradigmatic shift that these models allowed in understanding intelligence. Instead of measuring intelligence by the external behaviour, in line with the Chinese Room's idea (Searle, 1980), the analysis must be internal, as connectionist models permit inspecting their internal states (i.e. hidden unit activation patterns) of the acquired knowledge.

Thus, connectionism influenced explanations about cognitive phenomena, leading to a different thinking in this matter (Thomas & McClelland, 2008). There are several examples that can be given in the study of cognitive neuropsychology disorders or developmental disorders (Munakata & McClelland, 2003) but we will focus on cases with relevance for the scope of this thesis. One example where connectionism brought paradigmatic changes in conceptualizing the phenomena and its theories is memory. There is a shift on what is conceived as knowledge versus processing. On the classical perspective, memory is a place for information storage, whether random access memory or RAM or the hard disk that is a physically independent entity from the central processing unit or CPU which accesses memory information and operates upon it. In this computational view, the information that the hard disk contains, that can be seen as the long-term memory, is moved into the CPU, or the working memory, to be processed and the long term memories are discarded through RAM, a short-term memory buffer (Turing, 1950). In this paradigm, there is a clear physic distinction between what knowledge storage is and where it is processed.

On the contrary, in connectionist models, processing occurs via the propagation of activations throughout the network. Knowledge, in turn, is encoded in the network's weights between the processing units. knowledge, in this case, is not an entity that moves from one place to another but rather happens in the changes of connections, driven by experience and is attached to the struc-

ture of the model. Hence, information is processed in the same substrate where it is stored. Moreover, it is possible to establish a distinction between two types of knowledge representation, namely latent and active (Munakata, 1998). Latent knowledge representations correspond to information encoded in the connection weights that are built from prior accumulated experience. Active knowledge representations are present in the maintained activation states of the system, representing currently relevant information. These two types of knowledge representation establish a parallel with the long-term and short-term memory.

Another example that is relevant in the scope of this thesis is the contribution of connectionist models for the study of cognitive development (Plunkett & Sinha, 1992). Connectionist models, because they allow inspecting the model's representations which, in turn, are resulting of a learning algorithm that changes the patterns of connectivity as a function of experience, provide a means of observing the mechanisms from which development emerges. The possibility to observe different stages of development representations and the underlying mechanisms that drove the change is a fundamental property that makes these models more suited for studying cognitive development relative to symbolic, rule-based computational models (Elman et al., 1996). Contributes from connectionism in the study of cognitive development include developmental phenomena such as memory (Munakata, 2004) infant perceptual category development (Mareschal et al., 2000; Quinn et al., 1993), language acquisition (Bates et al., 2002; Christiansen & Chater, 2001), and reasoning in children (Gentner et al., 1995; Shrager & Siegler, 1998; Ahmad et al., 2002; Shultz & Sirois, 2008; Mareschal & Thomas, 2007). The issue of computational modelling of development will be addressed next.

3.1.3 Modelling development and learning

Development inevitably involves change. One of the most difficult issues in developmental psychology is the transition from one stage of functioning to another and the underlying mechanisms that produce that change. Without insights on the mechanisms that produce change, no comprehensive understanding of development is possible. In this regard, the problem is divided in (i) the *what*, that includes identifying such mechanisms and establishing the effects that the mechanisms produce and (ii) the *how* that specifies how the mechanisms operate (Siegler, 1989). It is fundamental to identify *what* develops in children and indeed this perspective concerned considerable attention (Sternberg, 1984). However, it is in the perspective of the *how* of development that computational modelling can build major contributions, that is, how knowledge is represented and how the transition is performed from one state of knowledge to the next (Shultz, 2003).

The central benefit brought by computational modelling in the study of cognitive development is the possibility of exploring the causal mechanisms, that is, focusing on how information is processed, rather than descriptive

approaches that focused on what are the infants' capabilities at any given age. Therefore, computational models provide a tool for understanding the information processing mechanisms and the processes involved in developmental change (Mareschal & Thomas, 2007). Moreover, computational models allow inspecting internal representations of the model, contributing for an account of how knowledge is represented.

The extent to which computational models capture development is not clear and it is an issue under debate. This depends indeed on the definition given to development in the first place and how it is differentiated from learning, if it is considered differentiable (see section 2.3.1). We identify two criteria for evaluating developmental tractability.

The first criteria has to do with the developmental transition from one stage into the next, that is related with (Klahr, 1984)'s definition, as stated: "It evaluates the extent to which the competing theories, which propose two different pairs of state descriptions for earlier and later competence, can be integrated with a transitional theory: one that can actually transform the early state into the later one. Regardless of the predictive power or elegance of a theory for a given state of knowledge, if there is no plausible mechanism that might have produced that state from some previous one, such a theory is seriously deficient." (Klahr, 1984, p.107)

The second criteria involves the constructivist perspective on development that goes further and claims that underlying the transition must also exist a structural qualitative change in the cognitive structure, leading to a previous stage with increasing and complex processing capacity (Mareschal & Shultz, 1996). "The constructivist view of cognitive development holds that children build new cognitive structures by using their current structures to interact with the environment. Such interaction with the environment forces adaptation to environmental constraints, and the adaptation of existing cognitive structures results in new, more powerful cognitive structures." (Shultz, 2003, p. 160)

Shultz & Marechal (1997) consider that connectionist and generative approaches can be reconciled and suggest directions in how to decide whether static or generative models are more appropriate to model development. They propose that connectionist networks are suited to be used in modelling basic universal cognitive skills because these abilities reflect regularities in the environments independently of cultural variations. Generative models, because they are able to construct their own architecture, hold the flexibility to apply quantification knowledge to a wide range of possible task domains. Thus, these models are more appropriated for high level cognitive skills that are built on top the initial core of abilities that begin to develop very early in infancy and that tend to vary significantly across the world.

The most applied computational techniques for the study of cognitive development are production systems, connectionist networks, dynamic systems, Bayesian inference and robotics (Shultz & Sirois, 2008). Production systems were a proposal for symbol manipulation in cognitive modelling, first

introduced by Newell (1973). These systems are long-term knowledge representations in the form of production rules, or condition-action pairs that determine actions or conclusions to be taken. Production rules are symbolic expressions that contain constants and variables. Such rules can also be called productions because of its capability to produce new knowledge. This means that in these symbolic architectures, symbols, or the assumed constituents of abstract cognition, are taken as modelling primitives. Examples of architectures for this approach include ACT-R (Anderson, 1993), Soar (Newell, 1990) and C4.5 (Quinlan, 1993).

In contrast, in connectionist networks, symbolic behaviour emerges from the operation of sub-symbolic processing units. These models process information by the propagation of activation between simple processing units. Knowledge is stored in the strength of the connections among units. Learning, therefore, happens through the gradual adjustment of the strengths of these connections.

The most common neural learning algorithms that are applied to development include back-propagation and its variants, cascade-correlation and its variants, simple recurrent networks, encoder networks, auto-association, feature mapping and contrastive Hebbian learning (Shultz & Sirois, 2008).

Given the networks' capabilities to learn and self-organize, the connectionist approach to development has gained extended interest in mechanisms of cognitive transition (Elman et al., 1996). These models address development as a consequence of non-linearities in the multilayer networks that produce graded transitions between different expertise stages. Generative networks, a variant of connectionist models, claim, as connectionist models, that learning takes place through connection-weight adjustments and development, in turn, occurs via a qualitative change in the structure that involves the recruitment of new units to the hidden layer of the network.

Dynamic systems are, in short, differential equations that determine how a set of quantitative variables change concurrently and interdependently, continually over time (Schoner, 2008). Dynamical systems address cognition through a theoretical framework within which an embodied view of cognition can be formalized. Neural networks can be considered dynamical systems, when a change in a state depends in part on values of current state. This overlap happens in: (i) recurrent networks, where the update of the activation depends in part of the current activation values and; (ii) also in learning networks where the update of weights depends in part on current weight values.

Bayesian inference are probabilistic models where the Bayes' rule is used to perform posterior inferences. Bayesian models have been gaining ground in cognitive phenomena and in the application to developmental problems (Griffiths et al., 2008). This framework for probabilistic inference presents a general approach for understanding how problems of induction can be solved and possibly how they might be solved in human mind. Three major contributions for modelling human cognition can be pointed out in these models.

The first is that Bayesian models establish a connection between human cognition and the normative prescriptions of a theory of rational inductive inference. Secondly, they contribute for the communication between different fields such as statistics, machine learning and artificial intelligence. Finally, these models can untie some theoretical debates that exist between models that emphasize symbolic representations and deductive inference and models that emphasize continuous representations and statistical learning such as neural networks.

Finally, developmental robotics challenge developmental modelling by employing the computational models inside of a robot, functioning in real environments and in real time. In this computational approach to development, the algorithms are embodied, turning robots into instances of models from developmental sciences (Lungarella et al., 2003). The intersection that robotics build between different disciplines such as artificial intelligence, artificial life, robotics, developmental psychology, neuroscience, and philosophy results in an approach that claims that brain, body and environment are connected. Consequently, cognition emerges from having a body that mediates perception by interacting with and moving in the real world.

3.1.4 Models of music perception and cognition

Music is present in every human culture and in different human activities. However, it is a culture-dependent phenomenon that adopts different forms, habits, and predispositional patterns. The forms of musical expression change with time and geographic location or culture. This attribute of ubiquity and, simultaneously, diversity makes musical trait a rich field. At the same time, music processing requires a series of psychological mechanisms such as learning and memory, attention, syntactic processing and processing of meaning information, emotion, action and social cognition. All these attributes make music an ideal means to study human cognition and can contribute with new insights for the understanding of the human brain (Koelsch, 2012; Pearce & Rohrmeier, 2012)

Models of music cognition aim capturing human musical knowledge, by transferring it to computers and integration into intelligent programs. Consequently, musical thought can be conceived as based on computations, connecting perceptions and actions, and is treated as empirically observable and formalizable (Laske, 1988).

Applications to music modelling

Music cognition is a very complex phenomenon and its exploration through a computational perspective is still underdeveloped. There is a lack of models that investigate music perception as a whole phenomenon or developmental related aspects such as how certain musical cognitive capabilities or predispositions develop from an initial infancy stage to a more adult-like mature

stage. The more common approaches to the problem take the scientific reductionism and aim to understand one element of the problem while holding the others fixed. There is a considerable amount of work done following this perspective, modelling music processing tasks (Purwins et al., 2008). Generally, computational models in music research can be distinguished between two approaches: (i) one that aims to model music knowledge, originating from music theory and; (ii) other that aims to formalize and understand the mental processes involved in music cognition (Honing, 2006). These models are specifically devoted to aspects of rhythm, melody and tonality perception. We will give a short overview on rhythm and melody which fit the scope of this thesis.

Generally, rhythm perception models focus on specific topics such as pulse finding, rhythm grouping and categorization, even though a comprehensive approach is still lacking (Purwins et al., 2008). In pulse-finding models, different approaches have been suggested. In rule-based models, pulse is suggested based on the regularities of the first audio events and then this hypothetical pulse is extended to the upcoming temporal patterns. Examples of this approach are Desain & Honing (1999) and Steedman (1977). Another account for modelling pulse finding is given by the idea of inner clocks which activation is induced by the perception of rhythm patterns (Povel & Essens, 1985). The internal clock is a regular pulse that is regulated to match the perceived rhythmic patterns as accurate as possible. Oscillator models address pulse in a signal processing perspective. Gasser et al. (1999) proposes an adaptive oscillator model that adjusts their frequency and phase to a sequence of input pulses. This model intends to model the perception of variable metrical timing patterns such as in music and speech. This system composed by a network of coupled oscillators is thus responsive with metrical structure stimulation, dealing with variation and reveals specific preferences based on experience. Todd (1994) proposes the primal sketch pulse detector, inspired by the theory of edge detection in vision (Marr, 1982). This multi-scale model of rhythm perception demonstrates to be successful for both music and speech signals. The output of the algorithm produces a "rhythmogram" that illustrates the input prosodic structure that can be interpreted as the rhythm grouping representation of the input in terms of Lerdahl & Jackendoff (1983)'s generative theory of tonal music. Finally, the contribution of robotics stresses the importance of the body for rhythm perception. Honing (2005) combines rhythm categorization with a kinetic model to address global tempo, note density and rhythmic structure.

Rhythmic grouping modelling is of special interest for this dissertation. In Chapter 6 we address the development of rhythmic representations and predispositions that consequently have influence in rhythm grouping. Therefore, the implementation of the computational model can somehow be related with this category of models.

Rhythmic grouping modelling has been addressed in less extent than

pulse finding. Unlike in pulse finding, where the perceiving is usually consensual, reinforced by notation and, hence, the analysis is clear, in grouping the correct analysis is much less certain. In grouping, human perceiving is much more ambiguous and vague than in pulse. Notwithstanding, there is enough agreement about grouping to make its modelling valuable (Temperley, 2004). Tenney & Polansky (1980) gave an important step in the computational study of grouping. Their algorithm is based on two grouping factors that derive from Gestalt principles, that is, proximity and similarity. The algorithm looks for smallest-level groups through the identification of local maxima in interval values. A different computational approach is developed by Baker (1989). In this approach that is intended for tonal music, a harmonic analysis is done to produce a hierarchical structure based on phrase structure rules similar to Lerdahl & Jackendoff (1983)'s approach. Mentioned before in pulse finding, Todd (1994)'s primal sketch is also a grouping model, where the input is analysed and peaks of energy are identified. These peaks lead to a hierarchical grouping structure. Finally, Desain & Honing (1991) present a connectionist model for rhythm categorization. Their numerical model considers consecutive durations and the internal representations of the relation between these durations converge to simple integer ratios. Extensive information on rhythm computational models can be found in and Gouyon (2005).

3.2 Computational modelling as a methodology

In this section we provide an overview of the methodology taken in this research, that is, computational modelling. For that, we will make a short description of the approach taken in computer simulations and a summary on the reasons that motivated the choice for this methodology. Apart from the benefits that computational modelling might represent, it is fundamental to be aware of the potential problems that one might face when dealing with this methodology. Therefore, we will cover very briefly these possible traps.

Music cognition is a complex subject and its study or understanding involves the contribution and crossing of many disciplines. Computational modelling is a territory where all the disciplines can be integrated and formalized into a computational model and connections can be built between the different disciplines. Moreover, computational modelling allows explaining musical thinking in the same way as everything else in science: by reducing a complex phenomenon as music cognition is into simpler components (O'Reilly & Munataka, 2000). Much as reductionism, cognitive computational models entail simplifying and identifying components that are mostly based on the physical substrate of the human cognition, that is, the brain. But complex system, in general, cannot be understood as a simple extrapolation from the properties of its elementary components. In this point, the computational approach enables combining the component pieces to recon-

struct the larger phenomenon. This step would be hardly achieved by using verbal models since it would be hard to use verbal arguments to reconstruct human cognition. This leads to the essential opportunity to observe interaction among the components. Hence, we get a complementary approach for analysing elements of a problem by dissecting a system to understand its essential elements and a later stage where these elements are "synthesised" or combined for understanding their interactions. Through the possibility of testing computer models, it is possible to observe non-expected emergent phenomena that could not be obviously present in the individual behaviour of the elements that arise from the interaction between the components. For this reason, computational modelling is a complete tool because it allows contemplating different kinds of explanation at different levels of description (microscopic and macroscopic level) that are linked into a cohesive whole (Marr, 1982).

For the computational models, we have adopted a connectionist approach (see Chapter 6). Connectionist models establish a detailed connection between biology and cognition in a way that is consistent with many established computational principles, since from a neuron-like structure high-level cognitive behaviours emerge (Shultz, 2003). Connectionist models allow focusing on computational processing of information in the brain and, at the same time, through that processing, observe and study high-level cognitive behaviours. This way, these models provide an account of the representations that underlie performance on a task that also incorporates a mechanism for the change of that representation. Not only computer models require being explicit about knowledge because they are implemented theories and data in terms of symbolic expressions but the sort of implementation, that is, distributed processing implementation, also leads to a simultaneous multiple level of analysis. Next, we summarize some of the advantages we found relevant in computer modelling approach for studying cognitive phenomena.

- **Explicitness.** The process of implementing a computer model forces to be precise about the assumptions taken and about how the relevant mechanisms work, avoiding some misunderstanding problems often related with verbal theories such as possible inexact arguments. Representations, symbols and variables must have an exact definition to allow implementation. This leads to confronting aspects of the problem that one might have otherwise ignored or considered to be irrelevant. Consequently, the algorithmic specificity results into detail and conceptual precision that implementing compels.

- **Contribution to a more explicative rather that descriptive perspective on cognitive phenomena.** Understanding the human mind only by observation of the human behaviour can be limited. Processes and mechanisms cannot be understood purely strictly on the basis of behavioural experimentation. They account on a superficial perspective of the features of human behaviour that are an emergent

result of the inner working processes and mechanisms that operate in the human mind. In this regard, computer models bring out an opportunity to complement this approach, and aim to explicate the intrinsic parameters and algorithms of the human mind. Therefore, computer models can contribute to a causal mechanistic understanding of cognitive phenomena, providing sources of insights into behaviour and explanations in the perspective the functioning of the underlying mechanisms. This way, computer models contribute to a transformation of the study of cognitive phenomena from a descriptive science into a explanatory science (Mareschal & Thomas, 2006).

- **Models' testability.** Computer models are implemented cognitive structures on a computer program that can be manipulated and tested by running simulations, enabling to explore the underlying mechanisms of cognitive phenomena. Simulations are opportunities for exploring different possibilities of details of a cognitive process and, furthermore, developing future theories. Moreover, models that implement a theory offer a means of testing internal self-consistency of that theory. The errors and failures that arise when implementing that theory will lead to a re-evaluation of the theory. This also can let emerge unexpected implications of the theory that, from the complexity of its nature, gives origin to interaction between its components.

- **Computational models are complementary** to empirical experiments and experimental data gathering, creating constraints on the direction of future empirical research.

- **Complexity.** Computer models can deal with complexity in the way that verbal models cannot, producing satisfactory explanations throughout its implementation. Computational models can handle complexity across multiple levels of analysis, allowing data across these levels to be integrated and related to each other.

- **Control.** Computer models allow control in the sense that many more variables can be controlled and with much more precision than in real systems. This allows exploring causal roles of different components in ways that would otherwise be impossible.

3.2.1 Problems with computational models

Although the advantages that cognitive computational modelling might have, there are drawbacks that one must be aware of, avoiding possible pitfalls when following this methodology. One common critic made to computational models is that they "can do anything". It is a very frequent assumed that, by changing and adjusting parameters, models can yield any desired output. It is true that computer models can have many degrees of freedom in their architecture. In particular, neural networks have numerous

parameters that determine the adaptation of weights between units. It is tempting to think that with so many parameters, fitting any behavioural phenomena might be simple and, therefore, not worthy. Other argument that supports this view is that different models can provide a reasonable explanation for one same given phenomenon. It is called as the indeterminacy problem (O'Reilly & Munataka, 2000). Relatively to the first argument, it must be present that most of the parameters in parallel processing models are not random or even adjustable by the modeller but rather determined by principled learning mechanisms. Due to the indeterminacy problem, this perspective on computer models can be outwit with an approach of striking an exhaustive testing to the model that although might be applied to a wider range of data, it is also applied in a greater detail on each task and properties of the learning process that the model must perform. This kind of approach avoids much likely that two different models can fit all the data (O'Reilly & Munataka, 2000).

Other risk that one might be exposed to in the computer modelling approach is when the assumptions taken for building the model lead to an implementation that is so simple that does not capture the relevant aspects of the phenomenon to be modelled and thus its validity becomes questionable. Building a computational model, because of the level of explicitness it requires, inevitably involves reducing and simplifying complex phenomenon. Indeed the assumptions that necessarily have to be made can be wrong. However, simplification can be beneficial for the model if the details that were omitted in the implementation are irrelevant and do not influence the results.

The opposite to over simplifications can also be a problem, as when models are too complex. When this happens, models and its behaviours become too difficult to understand and do not add any account on human behaviour because there are too much details interacting and influencing the results. This problem can be attenuated if the model is faced as an instantiation of wider principles rather than an end unto itself. These critical principles of the model must be clearly identified and articulated with the model's behaviour, demonstrating the relative insignificance of the other aspects that are excluded.

3.2.2 Computer simulations

In our approach, the integration of different disciplines is materialized in computer simulations. In a computer simulation there is a given cognitive phenomenon that we aim to study, that can also be named as the target. This target is mostly a dynamic entity, in the sense that it changes over time and reacts to its environment. The computer model creates an abstract specification of the target phenomenon that, in principle, should be simpler to study than the target itself. The simulation of the model is the means for

an analysis where the similarity between the predictions generated by the model and collected experimental data are evaluated (see Figure 3.1).

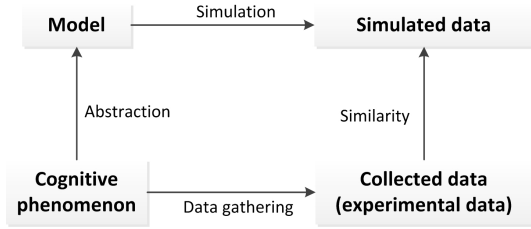

Figure 3.1: Simulation as a method, based on Gilbert & Troitzsch (2005).

However, given the dynamic nature of the target and, consequently, the need for dynamic non-linear models, not always the outcome of the model is entirely what was predicted in advance. It is not obvious to predict what the consequences of a model are, especially when dealing with complex phenomena. Therefore, simulations are crucial for testing the models. Simulations are an opportunity to test a computer model by running careful experiments and observing their behaviour under controlled conditions. This way, computer simulations can be seen as parallel to experimental methodology, where experiments are performed on a different sort of model's entities pool, instead of human subjects. Furthermore, computer models, because their internal representations can be inspected and analysed, gain a conceptual function in allowing the possibility to explain how the model solved some sort of problem. However, there must be precaution in assuming that the solution that the model presents for some mechanism of an observed behaviour is necessarily the same as the one that happens in humans. The plausibility of a model must be taken carefully since there is never a guaranty that the model is an accurate reflection of the human. Computer models are metaphors of the phenomenon aimed to be simulated and must be regarded as a resemblance to the real system they claim to model. On the other hand, because of the degree of unpredictability of models that can exhibit unexpected behaviours in simulations, models must be faced as rich sources of new hypothesis that can yield suggestions on new experiments to perform empirically.

Next, we enumerate the methodological procedures followed when building a computational model and testing it through simulations (Plunkett & Elman, 1997; Gilbert & Troitzsch, 2005).

3.2.3 Methodological procedures

"The process of model development within cognitive neuroscience is an exploration: a search for the key principles that the models must embody, for the most direct and succinct way of capturing these principles, and for a clear understanding of how and why the model gives rise to the phenomena

that we see exhibited in its behaviour." (McClelland, J. L. in (O'Reilly & Munataka, 2000, p. xxii))

As stated by McClelland, when building computational models and testing them through simulations, there are three items that must be defined right from the beginning: (i) a well defined hypothesis; (ii) a design for testing the hypothesis and; (iii) a plan for how to evaluate the result (Plunkett & Elman, 1997). However, these three main stages in building a computer model can be scrutinized into more detailed steps, as we do next:

- **Define a question** which will be the aim of the research to resolve. This step involves a clearly articulated problem and a well-defined hypothesis that is aimed to be tested. This first step is of extreme importance because it will influence the course of the rest steps. The question that is initially defined will influence the type of tests that will be performed in the simulation and, at the same time, the tests are limited by the design of the simulation.

- **Gather empirical data.** This includes data such as observations on the cognitive phenomenon that allow projecting the model and making decisions on the assumptions taken when designing the model and also data for constituting a training set. It is from this training set that the model will learn. Thus, the nature of training data is an issue of extreme importance, in terms of quantity and quality. There must be enough instances in the set and they must be as ecological and diverse as possible so that there is not the risk of extracting spurious generalizations.

- **Design the model**

  - Delimit a question to which the model should answer
  - Define clearly the assumptions taken when designing the model which state what is left out and what is included. This step involves making a set of choices to gain insight into the model. The options should be taken in order to achieve abstract and simplified models that better capture cognitive processes. The more is excluded, the more conceptual degree of the model and less complexity.
  - Define an algorithm
    - Motivate the nature of the stimulus representation used to feed the model. This deals with converting data into numerical codes and the decision of which representation to choose has consequences in the goals of the simulation.
    - Justify the use of the algorithm in the context of the empirical experiment

- **Define the task** that the model will perform. The task represents the behaviour that the model is trained to do. This can be, for example, learning to produce the correct output given an input. This implies the conceptualization of behaviour in terms of inputs and outputs.
- **Verification** of the correct implementation of the model by debugging the program
- **Test the model** by executing simulations
  - Expose the model to the training data set, training it on the task that was previously defined.
- **Evaluation** of the performance of the model on the simulation. This analysis is made through the validation of the model, by certifying that the behaviour of the model corresponds to the behaviour of the cognitive phenomenon, or its predictive and generalization power (Purwins et al., 2008). If there is some correspondence between what the simulation of the model reproduces and the observed data, the simulation can possibly represent a plausible model of the processes that led to the observed behavioural data. This can be done by measuring three parameters:
  - Individual pattern error that represents the difference between the output result of the model and the target output that the model should produce.
  - Global error is the averaged result of the overall individual pattern errors. In general, as learning progresses, this error declines.
  - Analysing internal representations. Techniques for doing so can involve hierarchical clustering of hidden unit activations that mean the measure of inter-pattern Euclidean distance. Inputs that are considered as similar by the model will produce similar internal representations, hence closer Euclidean distances (Plunkett & Elman, 1997). Other options resolve the previous approaches' limitation which is looking at the space directly, by visualising the hidden unit activation patterns. This can be achieved using principal component analysis together with projection pursuit or using the "Hinton diagrams" (Hinton, 1986), that involves looking at activation conjunction with the actual weights.
- **Understand the mechanisms** that caused the performance of the model. It is not sufficient an empirical validation, a theoretical and computational analysis are equally important for better understand models and the phenomena they are representing. This involves inferring from the data and thinking about the mechanisms that underlie

the production of the results. This process can lead to alternative accounts for the patterns of data. Then, a decision must be made for which of the several alternative proposals provides the "right" account.

4

Pitch Representation in Infancy

Summary

This chapter addresses pitch representation and its correlation with development. We approach this issue using computational tools. This way, we have carried out computational simulations that follow the setup and data for validation from behavioural experiments performed by Saffran & Griepentrog (2001). This research was developed in collaboration with Amaury Hazan, Perfecto Herrera and Hendrik Purwins (Salselas et al., 2008).

The computational model is supported in feed-forward neural networks and has been previously tested in computational simulations that involved solving tone sequence learning tasks in a framework that simulates forced-choice task experiments (Hazan et al., 2008). In the simulations, age is manipulated using different encoding types of the stimuli.

In the first simulation we test the models ability to perform a discrimination task based on absolute pitch cues, according their pitch encoding type. The second simulation represents a counterpart experiment, wherein the ability to perform discrimination tasks based on relative pitch cues is tested, according to the models pitch encoding type.

We have observed, through the simulations, a parallel between learning and the type of pitch information being used, where the type of encoding that was being used influenced the capacity of the model to perform the task correctly and learn to discriminate between music grammars. Moreover, the computational simulations' results support Saffran & Griepentrog (2001) hypothesis in the sense that infants may begin life with the capacity to represent absolute pitch and the relative pitch representation is developed later. Accordingly, the results achieved in the computational simulations were coherent with the findings reported in the behavioural experiments, validating the model and the encoding options taken. The simulation results revealed the model suitable for the simulation of absolute versus relative pitch representation and perceptual learning in infants and adults using a sequence learning task and added further validation to the model.

4.1 Pitch perception and representation in early development

Pitch is a fundamental perceptual attribute of sound and its detection is indispensable for encoding and memory storage of melody in music as well as in prosody in speech (McDermott & Oxenham, 2008; Trainor & Desjardins, 2002). Infants have the ability to categorize and discriminate single complex tones on the basis of pitch. Indeed, infants' resolution of frequency is finer than that required for musical purposes (Clarkson & Clifton, 1984). Five to 8 month old infants were tested for their thresholds for frequency differentiation and at a standard of 1000 Hz, presented at 70 dB, infants showed thresholds averaged at 21.6 Hz (Olsho et al., 1982). Moreover, infants' capacity for discriminating pitch depends on the spectral content of the sound, with experimental results showing that infants' performance in a pitch perception task deteriorates as the number of harmonics in a tonal complex decreases (Clarkson et al., 1996). The effect of number of harmonics held both for sound containing the fundamental frequency and for sounds lacking energy at that frequency. Experimental results have also shown that infants in early development discriminate more readily high pitches and also have a preference for this type of pitch (Werner & VandenBos, 1993). Maturation of the child's ability to discriminate frequencies may only reach adult levels until 7 years of age (Thompson et al., 1999).

In the process of learning about sequences of melodic tones or vowels and, consequently, in order to represent pitch, two main types of pitch cues can be used, namely absolute and relative pitch information. In absolute pitch representation, the cognitive processing of pitch is done independent of its relation to other pitch values, without any point of reference. In relative pitch representation, it is required a relational processing where distances between pitches must be considered and thus, there is no information about the specific fundamental frequencies (Levitin & Rogers, 2005).

It has been proposed (Mandler, 1992) that infants' attention is first more focused on absolute perceptual properties of a stimulus and then, progressively, becomes more abstract and, thus, attentive to structural relationships existent between different stimuli, similar to adult thinking. This substantially different way of processing information occurs, thus, through the emergence of qualitatively different modes of thought. In other words, there is a transition from unidimensional to multidimensional thinking (Siegler, 1996).

A transition also occurs in the auditory information processing. Younger infants, during a pre-conceptual period, focus on immediate perceptual dimensions of musical stimuli, such as pitch and timbre. In a later stage, as children develop and conceptual thinking increases, they learn more complex perceptual activities such as comparisons, transpositions and anticipation. This way, children focus on organization within the stimuli, such as rhythmic and melodic patterns, or contour, tonality and harmony (Sergeant & Roche, 1973). In experimental research, 3-year-old children showed greater tendency

to accurate representation of the pitch levels at which they had perceived the stimuli. Six-year-old children, in contrast, showed less concern for pitch level but more for melodic shape and sense of tonality (Sergeant & Roche, 1973).

The developmental shift in focus from absolute to relative features might also take place specifically in pitch representation. Accordingly, Saffran et al. (2005) have investigated the hypothesis that infants show developmental shifts in the focus on absolute or relative pitch information and that they have the capacity to encode pitches of sounds in an absolute way, independently of its relation to other sounds. In their experiments, results show that infants preferentially represent absolute pitches whereas adults preferentially represent relative pitches. Saffran et al. (2005) argue that infants use absolute pitch representation as a basic strategy for encoding auditory information, although relative pitch is also available. Accordingly, they hypothesize that infants may begin processing pitch in an absolute representation. This representation would be done via tonotopic frequency maps in the auditory cortex. Pitch contour, consequently, would be represented as a domain-general coding of up-down pitch change. Authors argue that absolute pitch processing is easy for an inexperienced brain as it is less computationally complex than relative pitch usage. Relative pitch information might be already available at this stage of development, but may be more complex to compute as it requires the detection of exact distances between pitches and the contrasting of multiple absolute pitch levels. However, in a later stage, through development, infants learn that relative pitch representation is more effective than absolute pitch representation and begin using it preferentially. Furthermore, speech perception requires the detection of relative distances between formants that contain acoustic information necessary for the recognition of phonemes and distinction between consonants. Hence, the use of relative pitch processing becomes essential.

Pitch perception and representation is also dependent on language experience (Cangelosi, 2005). Language experience may influence basic auditory processes such as pure tone perception at the level of auditory cortex. According to Deutsch et al. (2004), pitch representation is different in individuals that speak tone languages (Mandarin, Cantonese, Thai, and Vietnamese) from individuals that speak intonated languages such as English. In tone languages, words take different lexical meaning depending on pitch heights and also on their pitch contour. This may mean that in a critical period of development, tone languages' learners learn to associate words with pitches and pitch contours. Krishnan et al. (2005) tested the accuracy of pitch tracking and pitch strength in native speakers of Chinese Mandarin and English, measuring activity within the rostral brainstem. Chinese group exhibited stronger pitch representation and smoother pitch tracking than the English group. Researchers hypothesize that language experience may induce plasticity at the brainstem level. Depending on speech input, these adaptive neural mechanisms may influence pitch and pitch contour processing.

It is not clear how an absolute-pitch system develops into a relative-pitch processing one. Therefore, it becomes relevant to understand how this development occurs and, additionally, if it is somehow related to specific structural or training properties in the environment that could enhance, block or make possible such evolution. In order to explore developmental changes triggered by learning, in the types of perceptual information detected by the mechanisms underlying auditory learning, we have performed a computational simulation that follows a behavioural experiment (Saffran & Griepentrog, 2001). As this experiment provided the conducting thread for our simulation, we will subsequently proceed to its detailed description.

4.1.1 Absolute and relative pitch representation: experimental evidences

Saffran & Griepentrog (2001) ran two different experiments, where 8-month-old infants were examined in the use of absolute and relative pitch cues in a tone-sequence statistical learning task. In the first experiment, subjects were confronted with a learning problem that could only be solved if tones were represented by its absolute pitches. The second experiment is the counterpart design of experiment 1, where relative pitch pair statistics are the only available cues for solving the learning problem. A third experiment was performed, where adults were tested on the same statistical learning tasks used in the infant experiments, providing a cross-age comparison. For matter of simplicity, we will refer to two types of experiments, performed either with infants or adults, namely Experiment 1 and Experiment 2. The two types of experiments will be explored next.

Experiment 1: absolute pitch cues usage in a statistical learning problem

In this experiment, subjects were exposed to a continuous, unsegmented sequence of tones that served as a brief learning experience. The tone sequence was constructed by concatenating out of four tone words (see Table 4.1), with the stipulation that the same word wouldn't occur twice in a row, forming a 3 minute stream. Moreover, tone words did not resemble any paradigmatic melodic fragment or follow the rules of standard western-tradition musical composition.

Part-words (F C C♯ and D♯ G♯ A♯) are created by joining the final tone of one word to the first two tones of another word, spanning word boundaries. Words and part-words contain identical interval sequences. This means that part-words contain novel absolute pitch cues (combination of tones) but familiar relative pitch cues (intervals between tones). After the familiarization, a test period followed where subjects were confronted with pairs of words containing identical relative pitch sequences (see Table 4.1). The only information available for discrimination of words was absolute pitch cues.

Table 4.1: Tone words and test words used in the experiments by Saffran & Griepentrog (2001) in Experiment 1.

	Absolute Pitches	Relative Pitches
Tone words	G♯ A♯ F	M2↑ P4↓
	C C♯ D♯	M2↑ M2↑
	B F♯ G	P4↓ m2↑
	A D E	P4↑ M2↑
Test Words	B F♯ G	P4↓ m2↑
	A D E	P4↑ M2↑
	F C C♯	P4↓ m2↑
	D♯ G♯ A♯	P4↑ M2↑

The assessment of subjects' preferences was done by means of preferential listening methodology in the case of infants and forced-choice task methodology in the case of adults. In the preferential listening methodology, following the exposure phase, infants' listening preferences for words versus part-words are assessed by measuring the looking time at the source of the stimuli. For the forced-choice task, after the exposure phase, adults had to indicate the most familiar item between a word and a part-word (See Figure 4.1).

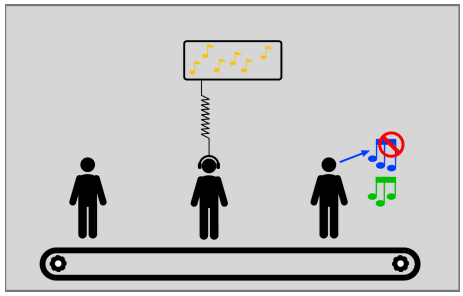

Figure 4.1: Methodology followed by Saffran & Griepentrog (2001) with their subjects.

Experiment 2: relative pitch cues usage in a statistical learning problem

Experiment 2 represents a contrasting test of the previous one performed. In this experiment, the methodology used with the subjects was the same but the stimuli used were different (see Table 4.2). In this case, test words

contained novel relative pitch cues but familiar absolute pitch cues. Consequently, discrimination was possible only on the basis of relative familiarity of relative pitch cues but not on the absolute pitch cues.

Table 4.2: Tone words and test words used in the experiments by Saffran & Griepentrog (2001) in Experiment 2.

	Absolute Pitches	Relative Pitches
Tone words	A♯ D♯ G	P5↓ M3↑
	F♯ D A	M3↓ P5↑
	G♯ A♯ F	M2↑ P4↓
	C♯ F♯ E	P4↑ M2↓
Test Words	A♯ D♯ G	P5↓ M3↑
	F♯ D A	M3↓ P5↑
	G♯ A♯ D♯	M2↑ P5↓
	C♯ F♯ D	P4↑ M3↓

Part-words (G♯ A♯ D♯ and C♯ F♯ D) consist of parts of two words rather that a sequence spanning a word boundary. Part-words contain novel relative pitch pairs (intervals between tones) but familiar absolute pitch pairs (combination of tones).

Behavioural experiments' results

The results indicated that adults and 8-month-old infants do not show the same pattern of learning performance given an identical set of tone sequence stimuli as input and that they based their discriminations on different types of pitch information. Infants succeeded at discriminating words from test words only for the contrast based on absolute pitch cues (Experiment 1) and failed to discriminate based on the relative pitch contrasts (Experiment 2). Adults showed the opposite pattern: successful discrimination based on Relative Pitch contrasts (Experiment 2) and no discrimination based on Absolute Pitch contrasts (Experiment 1).

Notwithstanding, Saffran & Griepentrog (2001) argue that the results obtained do not mean that infants can detect and use relative pitch information in tone sequence learning, or that adults do not retain residual absolute pitch abilities. The results show that given a sequential learning task, in which both relative and absolute pitch cues were available for discriminating words, infants rely more heavily on absolute pitch cues and adults rely more heavily on relative pitch cues. Possible explanations for the results obtained are pointed such as the atonal structure of the stimuli and the lack of musical structure on the stimuli.

4.2 Computer simulation goals

In this research step we aim to explore correlations between learning and pitch representation. Saffran & Griepentrog (2001) behavioural experiments concern early pitch perception and representation and, additionally, they approach this issue using a cross-age, developmental perspective. For the referred reasons, after a thorough analysis of Saffran & Griepentrog (2001) behavioural experiments, it becomes clear that they represent an excellent means for exploring pitch representation (absolute vs relative) and its relations with development. Therefore, the purpose of this research step is to investigate how learning (or exposure) influences the development of relative pitch representation given that infants might be initially equipped with absolute pitch representation. For that, we will perform computational simulations, based on these empirical experiments, applying different data encoding. This way, the simulations are a means for exploring absolute pitch representation in infancy and its development. At the same time, these simulations also correspond to a first step on computational modelling, allowing the practice and exploration of this methodology that we aim to pursue. The computational simulations are described hereafter.

4.3 Computational model overview

The computational model is supported in feed-forward neural networks (F-NN). F-NN's are artificial neural networks that process information in a parallel distributed mode and in its structure, the connections between the units never form a directed cycle. The network acquires knowledge through a learning process, and it is stored in the interconnection strengths or synaptic weight, resembling, this way, the brain functioning (Haykin, 2009). In sum, learning takes place by the continuous adaptation of the synaptic weights and bias levels, stimulated by the environment in which the network is embedded.

The connection weights of the network are updated applying back-propagation learning rule, implemented in an on-line setting. The type of learning is determined by the manner in which the parameter changes take place. In the specific case, learning is supervised, where the availability of a ground-truth set of training data is made up of N input-output examples.

The used back-propagation algorithm involves two phases: the forward phase, during which the parameters of the network (i.e. synaptic weights and bias levels) are fixed, the input signal is propagated through the network and, finally, the error signal is computed; and the backward phase, during which the error signal is propagated through the network in the backward direction and, hence, the adjustments are applied to the parameters of the network so as to minimize the error. Finally, back-propagation learning has been implemented in a sequential mode (or on-line mode, as referred before). This means that the adjustments made to the network's parameters are made on an example-by-example manner, in contrast with batch mode,

where adjustments are made at the end of each entire set of training examples (or epoch).

FNN's are suited for tasks such as next-event prediction, since they use past events as inputs and the next event to be predicted as output. This way, for the neural network to learn to predict the continuation of an encoded tone sequence, based on the tones observed so far, inputs that can be tones or intervals are presented to the input layer and successively transformed and propagated into successive layers via connection weights until activating the output layer, producing the prediction of the next event. Consequently, the model behaves so that the data in the input layer is successively transformed and propagated into the following layers, via connection weights, until the output layer is activated.

Moreover, the model had been tested previously in computational simulations that involved solving tone sequence learning tasks in a framework that simulates forced-choice tasks experiments (Hazan et al., 2008). For that, predictions of the model are compared with the actual data, for each word tone or interval (depending on the selected coding schema). The model performs the forced-choice task by selecting the word that possesses the lowest mismatch (between the model's predictions and the actual data).

4.3.1 Encoding

For the material used, three languages were created following the protocol used by Saffran & Griepentrog (2001):

- L_0: A training language to create training sequences;
- L_1: Language one containing words from the training language;
- L_2: Language two containing novel words

Words in L_0 are the same as the tone words in Saffran & Griepentrog (2001) as well as L_1 and L_2 are the test words. We have considered two options for the encoding of the sequences (see Figure 4.2).

- In *Pitch Class* encoding, each tone is encoded using a single pitch representation that contains meaning by itself, as in absolute pitch representation. In concrete, in this encoding, the FNN will receive 12 input units for representing a given pitch, and also the unit corresponding to the semitone is set to one
- In *Pitch Class Intervals*, there is a relational representation that considers intervals between elements and not its absolute value, as in relative pitch representation. For this case, the network receives 25 inputs that allow representing intervals ranging from -12 to +12 semitones. For an interval from one semitone to another, a specific unit is set to one

4.4 Simulations' setup

The computational simulations were performed following the behavioural experiments (Saffran & Griepentrog, 2001) that were previously described.

To carry out the simulations, we started from the assumption based on Saffran & Griepentrog (2001) experimental findings that, given a sequential learning task, in which both relative and absolute pitch cues were available for discriminating words, infants rely more heavily on absolute pitch cues and adults rely more heavily on relative pitch cues. Due to this, for the encoding of the material, we used *Pitch Class* encoding to simulate infant subjects and *Pitch Class Intervals* to simulate adult subjects. The different encodings involve the use of different number of input units. This way, age is manipulated with the type of encoding used (see Figure 4.2).

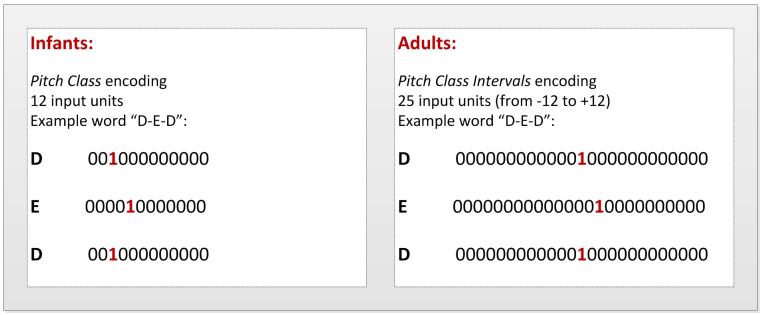

Figure 4.2: Encoding of the material: *Pitch Class* encoding was used to simulate infant subjects and *Pitch Class Intervals* to simulate adult subjects.

The experimental setup followed (see Figure 4.3) was in accordance with methods carried out by Saffran & Griepentrog (2001) with their subjects. In each run, one FNN, or *expectator*, is created, representing one subject, with random initial weights. Also, a learning sequence of encoded tones is generated using L0 and L1-L2 pairs of words for the forced choice tasks. The tone sequence was built according to the rules of the psychological experiment, in which tone words concatenated together in random order to create a 3 minute tone stream with the stipulation that the same tone word would never occur twice in a row.

Following, a pre-exposure forced choice task is performed in order to control any existing initial bias. Next, the *expectator* is trained, being exposed to the encoded tone stream previously generated and finally, one last forced choice task is performed in order to observe possible learning from the neural network.

Two simulations were carried out, correspondingly to the behavioural experiment aimed to be simulated. In simulation 1, the same tone words were used as in Table 4.1 and in simulation 2 as in Table 4.2. For each

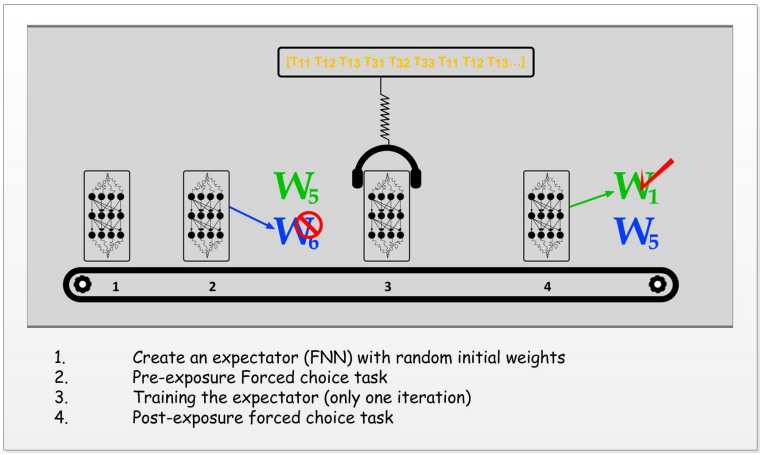

Figure 4.3: The experimental set-up followed for the simulations.

simulation, 50 runs were performed.

4.4.1 Simulation 1: absolute pitch cues usage in a statistical learning problem

This simulation aims to reproduce the behavioural experiment (see Experiment 1 in section 4.1.1) performed by Saffran & Griepentrog (2001). This way, the tone words used for this simulation are the ones shown in Table 4.1. Underlying the structure of these words is the fact that only absolute pitch contrasts are available for discrimination. Therefore, only an "infant" *expectator*, that is, a network trained using pitch class encoding, would be able to discriminate a familiar language L_1 from a novel language L_2. In the same way, an "adult" *expectator*, that is, a network trained using pitch class interval encoding, wouldn't be able to learn L_0 and, consequently, wouldn't discriminate L_1 from L_2. Figure 4.4 illustrates the results achieved in this simulation, using pitch class encoding, that is, for "infant" *expectators* and the tone words used by Saffran & Griepentrog (2001) in Experiment 1.

Figure 4.5 illustrates the results, using pitch class interval encoding, that is, for an "adult" *expectator* and the tone words used by Saffran & Griepentrog (2001) in Experiment 1.

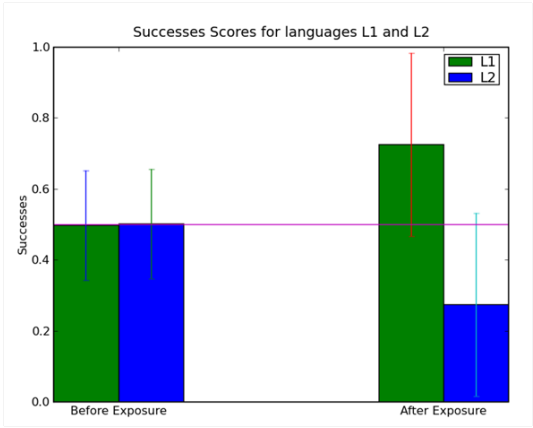

Figure 4.4: Results for "Infant" *Expectators* with **Pitch Class** encoding, before exposure and after exposure (50 runs), in simulation 1. Vertical axis represents percentage of successes.

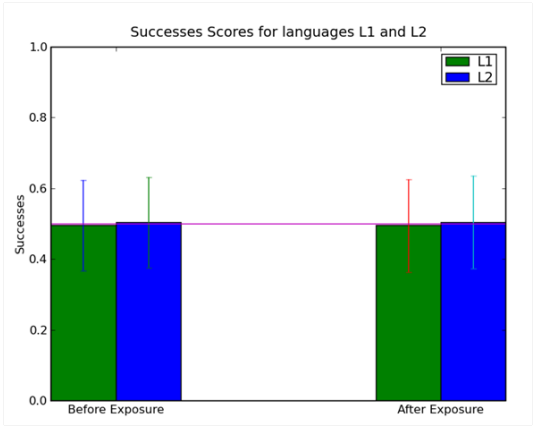

Figure 4.5: Results for "Adult" *Expectators* with **Pitch Class Interval** encoding, before and after exposure (50 runs), in simulation 1. Vertical axis represents percentage of successes.

4.4.2 Simulation 2: relative pitch cues usage in a statistical learning problem

This simulation aims to reproduce the behavioural experiment (see Experiment 2 in section 4.1.1) performed by Saffran & Griepentrog (2001). This way, the tone words used for this simulation are the ones shown in Table 4.2. Underlying the structure of these words is the fact that only relative pitch

contrasts are available for discrimination. Therefore, only an "adult" *expectator*, that is, a network trained using pitch class interval encoding, should be able to discriminate a familiar language L_1 from a novel language L_2. In the same way, an "infant" *expectator*, that is, a network trained using pitch class encoding, shouldn't be able to learn L_0 and, consequently, wouldn't discriminate L_1 from L_2. Figure 4.6 illustrates the results achieved in this simulation, using pitch class encoding, that is, for "infant" *expectators* and the tone words used by Saffran & Griepentrog (2001) in Experiment 2.

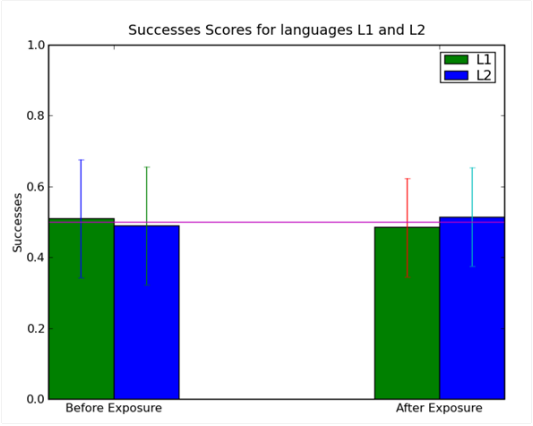

Figure 4.6: Results for "Infant" *Expectators* with **Pitch Class** encoding, before and after exposure (50 runs), in simulation 2. Vertical axis represents percentage of successes.

Figure 4.7 illustrates the results, using pitch class interval encoding, that is, for an "adult" *expectator* and the tone words used by Saffran & Griepentrog (2001) in Experiment 2.

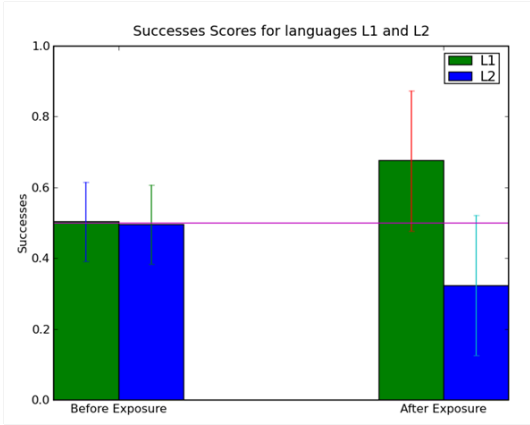

Figure 4.7: Results for "Adult" *Expectators* with **Pitch Class Interval** encoding, before and after exposure (50 runs), in simulation 2. Vertical axis represents percentage of successes.

4.5 Results and discussion

The model was successful in learning languages given different types of encodings. In simulation 1, it was possible to simulate the learning of L1 using Pitch Class encoding, when only absolute pitch contrasts were available for discrimination (see Figure 4.4). In Figure 4.5 it can be observed that the *Experimenters*, not being able to learn L_1 based on Pitch Class encoding, choose words randomly either before or after exposure.

In contrast, in simulation 2, the model was unable to learn L_1 given Pitch Class encoding (see Figure 4.6) but it was possible to simulate the learning of L1 using Pitch Class Interval encoding, when only relative pitch contrasts were available for discrimination (see Figure 4.7).

These results are coherent with the findings reported by Saffran & Griepentrog (2001) in their psychological experiments, showing a parallel between learning and the type of pitch information being used. Saffran & Griepentrog (2001) report that adults and infants based their discriminations on different types of pitch information. Infants succeeded at discriminating words from test words only for the contrast based on absolute pitch cues and failed to discriminate based on the relative pitch contrasts. Adults showed the opposite pattern: Successful discrimination based on relative pitch contrasts and no discrimination based on absolute pitch contrasts.

In the case of the computational simulation, learning after exposure of L1 in both cases of Experiment 1 and Experiment 2 depends on the type of pitch encoding used. Using Pitch Class encoding in order to simulate infant subjects, the system showed post-exposure learning scores when having available absolute pitch contrasts. In turn, using Pitch Class Interval encoding in order to simulate adult subjects, post-exposure learning scores were observed only when having available relative pitch contrasts.

This way, the model has shown suitable for the simulation of absolute versus relative pitch representation and perceptual learning in infants and adults using a sequence learning task by being able to yield different types of decisions, depending on the encoding utilized.

4.6 Concluding remarks

Throughout the simulations, we have observed that the type of encoding that was being used influenced the capacity of the model to perform the task correctly and learn to discriminate between languages. The simulations' results were coherent with the experimental ones. The encoding chosen for "infant" *Expectators* allowed the model to perform in conformity with behavioural experimental results obtained for infants. In turn, the encoding chosen for "adult" *Expectators* allowed the model to perform in conformity with behavioural experimental results obtained with adults. Thus, we conclude that manipulating age with the type of encoding was a correct assumption. Moreover, the simulations' results also represent additional validation to the model that had been previously tested in computational simulations (Hazan et al., 2008). The simulation results, this way, corroborate Saffran & Griepentrog (2001) conjecture about infants beginning life with the capacity to represent absolute pitch and the relative pitch representation is developed later.

Further than having models that successfully perform tasks using different types of encoding, it would be a challenge to explore the mechanisms that could lead one type of encoding evolving to the later one. Generative models could eventually be a means to study this evolving process. These models allow structural change in the model, leading to a previous stage with increasing and complex structure (see section 3.1.3). This characteristic adaptation would potentially permit the model to develop from one type of encoding to the next. The fact that models, using different encoding types, have different performances also points out that choosing the type of encoding is a matter of relevance. For this reason, the selection of the type of encoding to be used, when building a computational models of music, becomes extremely important.

Regarding the methodology used, we consider that computational modelling is a suitable means for studying cognitive phenomena, worthwhile to be explored in further research. As refereed before, language experience may in-

fluence basic auditory processes, among which pitch perception would be one of them. For this reason, language should be included in the next research stages, pursuing a comparative approach where both music and language are included. The comparative approach might contribute for exploring further influences between music and language and possible shared processing mechanisms.

5

Prosodic characterization in music and speech: exploring key factors during early development

Summary

In this chapter we pursue the hypothesis whether the infants' development of musical predispositions is influenced by the prosodic features that are present in the sonic environment of their culture. We consider that an important step towards the exploration of this hypothesis is to first understand what are the key elements for characterizing the prosodic environment of an infant. In this regard, in this chapter we aim to capture rhythmic and melodic patterning in the most salient auditory information to which infants are exposed, that is, speech and singing directed to infants.

We address this issue by exploring the acoustic features that best predict different classification problems. We built a database composed by infant-directed speech from two Portuguese variants (European vs Brazilian Portuguese) and infant-directed singing from the two cultures, comprising 977 tokens. These two Portuguese variants share the same lexicon and thus the prosodic differences between them would be the variable to focus on.

In the first machine learning experiments conducted, we aimed to automatically discriminate between language variants for speech and vocal songs in order to explore the acoustic properties that best differentiate the two Portuguese variants. Descriptors related with rhythm exhibited strong predictive ability for both speech and singing language variants' discrimination tasks, presenting different rhythmic patterning for each variant. Moreover, common features could be used by a classifier to discriminate speech and singing, indicating that the processing of speech and singing may share the analysis of the same stimulus properties. These results suggest that further exploration of music and language processing parallels and interactions should be taken. With respect to the experiment aiming to discriminate between interaction classes, pitch-related descriptors showed better performance.

We conclude that prosodic cues present in the surrounding sonic environment of an infant are rich sources of information not only to make distinctions between different communicative contexts through melodic cues,

but also to provide specific cues about the rhythmic identity of their mother tongue. Consequently, these rhythmic prosodic differences and the influence they might have in the development of the infant's musical representations will be further explored in Chapter 6.

5.1 Introduction

Early experience has a fundamental role in brain development. In this critical period, developmental processes are especially sensitive to environmental input, and the acquisition of adult level abilities in specific areas is dependent on the surrounding stimuli or the lack of it (Patel, 2008).

Among the auditory information to which infants are exposed, the most salient are speech and singing sounds. Parents and caregivers, across cultures, languages and musical systems, use a distinctive register for singing and speaking to their infants (Papoušek & Papoušek, 1991; Trehub et al., 1993). Regarding singing, caregivers usually use a special selection of music, consisting of lullabies and play songs. These are sung to infants in a particular style of singing that is different from the typically adult style (Trainor et al., 1997). These acoustic modifications in infant-directed singing attract the infant's attention and may be used by adults to regulate infant states and to communicate emotional information (Rock et al., 1999).

In infant-directed speech, also called motherese, there are acoustic adjustments in speech elements such as hyper-articulation, with more extreme vowel formant structure, higher mean pitch, wide pitch range, longer pauses and shorter phrases (Papoušek et al., 1987). In addition to engaging and maintaining the infant's attention, these distinctive modifications play an important role for indicating different communicative intentions to pre-verbal infants, such as to arouse or to soothe and to convey approval and prohibition (Fernald, 1993).

The meaning of the melodies present in maternal speech has been studied and the form of the melodic contours has been categorized according to contour shape (Fernald, 1989). Performing an acoustic analysis of utterances, prototypical contours were found for specific interaction classes (Papoušek et al., 1990). These prototypical shapes have been considered cross-linguistic universals (Papoušek & Papoušek, 1991).

From the perspective of a pre-verbal infant, music and speech may be not as differentiated as they are for older children and adults. They may be perceived as sound sequences that unfold in time, following patterns of rhythm, stress and melodic contours. Therefore, before the availability of verbal communication, the prosodic information present in speech and music domains such as melodic and rhythmic cues are primarily a communication system, a pre-linguistic system or a "prosodic protolanguage" (Masataka, 2009).

Culture-specific perceptual biases (such as sensitivity to language-specific

rhythms) emerge during infancy and may be acquired by being exposed to the speech and music of a particular culture. It is possible that the statistical information present in the sonic environment of infants shapes their preferences for certain contours (sequences of pitches and durational contrasts), and thus the exposure to speech and music with different prosodic characteristics could result in the development of different melodic representations. Comparing the rhythmic and melodic patterning in speech and music should shed some light on this issue. Additionally, a cross-varietal examination of prosodic differences may help to distinguish between generic features (that are shared and exploited in different cultures) and specific features of a given speech culture.

We have selected Brazilian and European Portuguese for pragmatic reasons. These two Portuguese variants share the same lexicon (verbal content) and thus the prosodic differences between them would be the variable to focus on. The conduct of this study will lead to further investigation in how prosodic patterning from each Portuguese variant may influence the infant's development of different melodic representations or predispositions in each culture.

The processing of speech and singing may require the use of the same perceptual processes and of similar cues such as durational (or rhythmic) and pitch patterning. Therefore, we also aim to explore if the same features are used to perform speech discrimination and singing discrimination tasks, in order to verify if the cognition of music and language share perceptual cues and computational characteristics during the pre-verbal period. Also, we aim to investigate if the features used to discriminate the variants of speech and singing are specific to this task or if they are also discriminative in a different condition, such as an interaction context discrimination task.

After a brief background review, we explain in section 5.2 how we gathered relevant samples of infant-directed speech and infant-directed singing, and how rhythmic and melodic features were extracted from them in order to devise and test different classification models based on task-related prosodic properties. In section 5.3, different classification experiments will be reported. Section 5.4 presents the discussion of the results obtained, and the last section presents our conclusions.

5.1.1 background

Prosody in both music and speech manipulate acoustic features to convey emotional expression and to provide segmentation and prominence cues to the listener. Speech prosody refers to speech properties that go beyond sequences of phonemes, syllables or words, that is, the supra-segmental properties of speech. These characteristics comprise controlled modulation of the voice pitch, stretching and shortening of segments and syllable durations, and intentional loudness fluctuations (Nooteboom, 1997).

Speech intonation or melody is related with speaker-controlled aspects of voice pitch variations in the course of an utterance. These pitch variations can have similar patterns, and thus languages can be organized as intonation languages, such as the Germanic, Romance and Japanese languages, or as tone languages, such as Chinese, in which words take different lexical meanings depending on pitch pattern (pitch heights and pitch contours). Although speech melody is perceived by listeners as a continuous streaming of pitches, in fact it is interrupted by the production of voiceless consonants such as /p/, /t/, /k/ that introduce silent intervals or pauses. Therefore, pitch is perceived in voiced pitch (quasi-periodic complex sounds) such as vowels.

Prosodic rhythmic properties are related to temporal aspects of speech and involve the patterning of strong beats or prominent units alternating with less prominent ones. The study of speech rhythm focuses on the organization of sound durations and its contrasts, that compose the temporal patterning of speech. Different factors contribute to the perception of these durational variations (Santen & Olive, 1989). However, the definition of the durational units, and thus, which duration units are more salient from a perceptual point of view, remains controversial. Furthermore, speech rhythm may be a consequence of the perception of time-specific events like beats, and not durational units.

In the study of prosody and language, different durational units have been considered. Vocalic intervals are defined by the section of speech between vowel onset and vowel offset. Consonant intervals or intervocalic intervals are defined as the section between consonant onset and consonant offset (Ramus et al., 1999). Other durational units have also been considered such as Inter-Stress Intervals (ISI) or the duration between two successive stresses, the duration of syllables, and the V-to-V durations (Barbosa, 2007) or intervals between successive vowel onsets, which are considered to be perceptually equivalent to syllable-sized durations.

Languages have been categorized into rhythm classes based on the notion of isochrony (Pike, 1945). These classes would typically be syllable-timed, stressed-timed and mora-timed languages. A contrasting approach is that languages would be organized in rhythm along a uniform continuum space rather than in cluster classes (Grabe & Low, 2002). European Portuguese and Brazilian Portuguese have been found to be clearly distinct in rhythm patterning (Frota & Vigario, 2001). European Portuguese is considered to have a mix of both stress and syllable-timing rhythm patterning while Brazilian Portuguese is considered to have a mix of syllable and mora-timing rhythm patterning. Thus, these two variants from the same language share the same words (lexical content) but differ in prosodic properties.

Infants are very sensitive to prosodic information. They can retain surface or performance characteristics of familiar melodies in long-term memory. These are said to contribute to the perception of the expressed emotional meaning. In particular, infants can remember specific details of tempo and

timbre of familiar melodies (Trainor et al., 2004). Prosodic cues are also fundamental for infants in speech domain. Infants primarily focus on acoustic features of speech such as prosodic information rather than phonetic or lexical information. Moreover, newborn infants are able to categorize different speech rhythms, as they discriminate their mother tongue from languages belonging to different standard rhythmic classes.

Infants can discriminate speech rhythm classes with a signal filtered at 400 Hz, which suggests that they probably rely on distinctions between vowels and consonants to accomplish the discrimination task (Mehler et al., 1996). These findings point to rhythm based discrimination by newborns (Nazzi & Ramus, 2003). Thus, prosodic features play an important role in the acquisition of both music and speech, as they provide information to segment continuous streams into meaningful units and to learn about their structures.

Music and language cognition and its interactions have been addressed with diverse scientific approaches. Some studies are oriented to explain cognitive phenomena, as it is the case of Patel et al. (2006), who studied language and music relations by quantitatively comparing rhythms and melodies of speech and of instrumental music. This study has shown that music (rhythms and melodies) reflects the prosody of a composer's native language. Also supporting the suggestion that musical rhythm of a particular culture may be related with the speech rhythm of that culture's language, Hannon (2009) demonstrated that subjects can classify instrumental songs composed in two languages that have different rhythmic prosody basing their decisions on rhythmic features only.

In a different approach, language and its rhythmic and melodic properties have been explored by looking forward to design automatic recognition systems such as automatic language identification, automatic emotion recognition in speech, and speech synthesis. In these artificial systems, speech is automatically segmented into rhythmic units (syllable, vowel, and consonant intervals). The temporal properties of these units are then computed and statistically modelled for the identification of different languages (Rouas et al., 2005). For segmentation, spectral information is extracted, consonants are identified as abrupt changes in the wave spectrum, and vowels are detected by locating sounds matching vocalic structure by means of spectral analysis of the signal (Pellegrino & Andre-Obrecht, 2000).

Galves et al. (2002) propose a different approach to segmentation which is based on the measure of sonority defined directly from the spectrogram of the signal. This means that two types of portions of the signal (sonorant and obstruency) are identified: sonorant parts exhibit regular patterns, and obstruency portions exhibit the opposite pattern, similarly to vowels and consonants. In automatic identification of emotional content in speech, features of the signal such as pitch (pitch range), intensity, voice quality and low-level properties such as spectral and cepstral features have been explored.

Slaney & McRoberts (2003) used pitch, broad spectral shapes and energy variations to automatically classify infant-directed speech into different communicative categories. To characterize the broad spectral shapes, they used mel-frequency cepstral coefficients (MFCC's). Automatic identification of emotional content in speech has also been applied to categorize different communicative intentions in infant-directed speech. For this task, supra-segmental features are examined such as statistical measures of fundamental frequency and properties of the fundamental frequency contour shape (Mahdhaoui et al., 2009; Katz et al., 2008).

In the present study, we will make use of computational techniques, linguistic and psychology knowledge with the purpose of understanding music and speech categorization by infants. Methods used to carry out this study will be described in the next section.

5.2 Methods

5.2.1 *Corpus*

For the construction of the audio database that served as a basis to our study we considered infant-directed speech and infant-directed singing from Brazilian Portuguese and European Portuguese. European Portuguese was taken from recordings captured for the purpose of this study. Brazilian Portuguese infant-directed speech and singing was compiled taking samples from the CHILDES database (MacWhinney, 2000), specifically from an audio database compiled to study rhythm acquisition (Santos, 2005) and from on-purpose captured audio. All audio signals considered were digital, stereo, 16 bit at 44100 Hz.

The recordings contain caregivers interacting with their healthy babies aged up to 18 months. During the recordings, caregivers were interacting with the babies at their home and in different contexts such as playing, feeding, bathing and putting them to bed. The materials contain spontaneous interactive speech and singing. The database is comprised by 23 adult caregivers, 9 Brazilian Portuguese subjects (2 male and 7 female) and 14 European Portuguese subjects (3 male and 11 female). For the singing materials, a subset of subjects is represented. For European Portuguese there are six singing subjects, and for Brazilian Portuguese there are five singing subjects. Each singing class contains 20 playsongs and 8 lullabies.

Subsequently, the audio from the recordings was cut into utterances that we refer to as interaction units. Four interaction classes were considered:

- Affection: a positive affect to provide comfort to the infant such as "Ohhh my sweet baby"
- Disapproval: a negative affect such as "No!! Don't do that!"

- Questioning: a more complex sound sequence such as "Would you like to have a cookie?"
- Singing: considering play songs and lullabies sung while interacting with the baby

These sounds were used as the instances for all the experiments reported in this paper, organized and grouped into different manners, as will be described. Instances gathered are summarized in Table 5.1.

Table 5.1: Organization of the instances gathered

Brazilian Portuguese		European Portuguese	
Affection	**151**	Affection	**162**
Disapproval	**150**	Disapproval	**150**
Question	**156**	Question	**152**
Singing	**28**	Singing	**28**

Utterances that were used to build the database were recorded in spontaneous interaction contexts. As such, the materials do not contain exactly equivalent text (sentences) for each variant. However, when recorded, subjects spoke the same language, Portuguese, and they were making use of the same word dictionary (lexicon). The database contains a sufficient number of instances (977) to ensure a variety of elements that can be considered comprehensive.

Because of the amount of instances collected, and because of the use of the same interaction contexts in both language variants, it is unlikely that a lexicon bias appears in the corpus. According to these considerations, we trust the database as being representative of the classes we try to model and compare, and thus we can generalize from these particular examples.

As infant-directed speech was recorded in the context of spontaneous interactions, it was very difficult to select portions of audio that belonged to a given interaction class and that were not mixed with background noise, such as, for example, babbling and noise from the baby's toys. For this reason, the amount of data (instances) is somehow limited. On the other hand, the data considered is spontaneous and it was collected from recordings of four different interaction contexts. Therefore, for its variety in content, the corpus can be considered representative.

5.2.2 Discrimination system model

Automatic segmentation method

For the segmentation of the durational units in the utterances, we used Prosogram (Mertens, 2004). The main purpose of Prosogram is to provide

a representation of intonation, considering that auditory perception of pitch variations depends on many factors other than $F0$ variation proper. Prosogram produces a representation that aims to capture the perceived pitch patterns of speech melody (a stylisation based on perceptual principles). Four perceptual transformations to which speech is subject are taken into account; specifically, segmentation into syllabic and vocalic nuclei, a threshold for the detection of pitch movement within a syllable or the glissando threshold, the differential glissando threshold (a threshold for the detection of a change in the slope of a pitch movement in a syllable) and temporal integration of $F0$ within a syllable. Figure 5.1 illustrates a pitch contour stylisation from Prosogram.

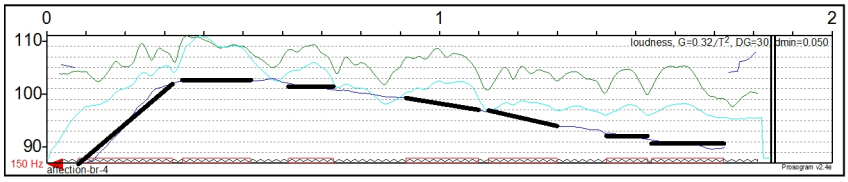

Figure 5.1: Illustration of the Prosogram of an affection instance ("hmmmm nham nham nham nham nham"). Horizontal axis represents time in seconds and the vertical axis shows semitones (relative to 1 Hz). Green line represents the intensity, blue line the fundamental frequency, and cyan the intensity of band-pass filtered speech.

Prosogram is a suitable tool for studying music and language (Patel et al., 2006; Patel, 2006) since the representation produced consists on level pitches and pitch glides. Hence, we have applied this method for speech and singing. We used Prosogram to extract, from the interaction units, vocalic intervals' onset and offset, intervocalic intervals' onset and offset, and pitch value within vocalic intervals.

This automatic segmentation algorithm does not require preliminary segmentation into sounds or syllables. It uses local peaks in the intensity of band-pass filtered speech, adjusted on the basis of intensity, to segment the signal. F0 detection range was set to 40 to 800 Hz, with a frame rate of 200 Hz. The glide threshold used was $0.32/T^2$ semitones/s, where T is the duration of a vocalic nucleus in seconds.

An evaluation to assess Prosogram's reliability for automatic segmentation was performed. We compared Prosogram's automatic detection of vowels against a ground-truth made with manual annotations. The *Vowel Error Rate* (VER) (Rouas et al., 2005; Ringeval & Chetouani, 2008) was used to evaluate Prosogram, as well as vowel onset and offset detection. VER is defined follows:

$$VER = 100.\frac{(N_{del} + N_{ins})}{N_{vow}}\% \tag{5.1}$$

where N_{del} is the number of vowels deleted or not detected, N_{ins} is the number of inserted vowels and N_{vow} is the reference number of vowels provided by manual annotation. We have manually annotated a subset of 96 instances from the materials (15 from each speech class and 3 from each singing class) that represent approximately 10% of the whole corpus. Table 5.2 shows the total number of vowels hand-labelled (Reference N_{vow}), detected by Prosogram (Detected), inserted (Inserted N_{ins}) and non-detected (Deleted N_{del}) and finally VER value. The VER value is considerably low when comparing with VER values obtained by Ringeval & Chetouani (2008).

Table 5.2: Prosogram's performance compared with hand labelling.

Reference N_{vow}	Detected	Inserted N_{ins}	Deleted N_{del}	VER
592	558 (94.26%)	15 (2.53%)	34 (5.74%)	8.27%

In order to complete the evaluation, we assessed Prosogram's detection of the onset and offset of vowels. We used a tolerance window of 25ms, which is approximately 10% of the annotated vowel average durations. We obtained 80% precision ($F-measure = 0.796$) for onset detection and 56.6% precision ($F - measure = 0.569$) for offset detection. Thus, Prosogram proved to be very helpful in providing a reliable automatic detection and saving a cumbersome hand-labelling task.

Durational units considered

The vowels' onset and offset obtained using Prosogram were used to compute three different durational units: vocalic intervals (V), consonant intervals (C), and V-to-V intervals.

Vocalic intervals where computed considering the section of speech between a vowel onset and a vowel offset. A vocalic interval may then contain more than one vowel and can span a syllable or word boundary. Consonant intervals or intervocalic intervals consist of portions of speech between vowel offset and vowel onset. We are considering these durational intervals with the assumption that infants can distinguish between vowels and consonants.

Ramus et al. (1999) argue that infants perform a crude segmentation of the speech stream which only distinguishes vocalic and non-vocalic portions, and classify different languages based on this contrast. In addition, in languages with rhythmic patterns close to stressed-timing such as European Portuguese, stress has a strong influence on vowel duration. Marking certain syllables within a word as more prominent than others leads to vowels consistently shorter or even absent, in contrast to Brazilian Portuguese where there is small contrast in the duration of adjacent syllables.

V-to-V durations were computed as the interval between successive vowel onsets (Barbosa & Bailly, 1994; Barbosa, 2007). V-to-V units are considered perceptually equivalent to syllable-sized durations, a fundamental unit for

speech perception (van Ooyen et al., 1997). It is relevant to consider here these durational units given that infants are responsive to syllable patterning and these units are particularly salient during the initial period of speech acquisition and processing, regardless of the language and rhythmic pattern of the stimuli (Bertoncini et al., 1995).

Extraction of descriptors

After computing the temporal measures just described, we proceeded to compute descriptors in order to capture melodic, temporal and accentual prosodic patterns of the speech and singing materials. Descriptors were computed separately for each instance. We have divided the descriptors into two categories: pitch-related and rhythm-related descriptors. A brief description of these descriptors follows.

(a) **Rhythm-related descriptors**: Normalised pairwise variability index (nPVI) was computed for the vocalic intervals and for the V-to-V intervals in order to measure the contrast between successive durations, which may reveal changes in vowel length within interaction units (Ling et al., 2000). Higher overall nPVI should occur in the European Portuguese variant, in which vowel reduction and consonant clustering are characteristic, leading to greater durational contrast.

For consonant intervals, raw pairwise variability index (rPVI) was computed. nPVI was not considered for this type of durations because it would normalize for language variant differences in syllable structure (Grabe & Low, 2002). Also, this descriptor could reflect consonant clustering due to potential vowel suppression in European Portuguese but not in Brazilian Portuguese.

Standard deviations were calculated for vocalic, consonant and V-to-V durations. Coefficients of variability (std/mean) were also computed for the three duration types in order to measure the variability of durations. These measures may not be directly relevant to the perception of rhythm but may reflect, as global statistics, the variability in syllable structure (Patel, 2008). Finally, speech time, the proportion of vocalic intervals in an interaction unit (%V) or the percentage of speech duration devoted to vowels, and speech rate (number of vocalic intervals per second) were also computed.

(b) **Pitch-related descriptors**: nPVI and coefficient of variability were computed for the median pitch of each vocalic interval in order to measure the contrast between pitch values and pitch variability, respectively. The lowest pitch value, highest pitch value, pitch range, mean and standard deviation pitch value for each interaction unit were also calculated. Finally, the percentage of vocalic intervals in which pitch is flat, rises, and falls were computed.

Additionally, descriptors related with the overall pitch contour were extracted aiming to capture pitch shape patterns. A polynomial regression was performed, using the median pitch values of each vocalic interval as points, in order to fit the pitch contour.

Next, kurtosis, skewness and variance were extracted from the pitch contour approximation previously calculated. Dividing this approximation curve into three equal portions, the slope of the beginning, middle and end of the curve was then calculated.

Attribute selection

In order to identify a group of relevant descriptors for class discrimination, we performed an attribute selection using the Correlation-based Feature subset Selection (CFS). The CFS algorithm (Witten & Frank, 2005) uses a correlation-based heuristic for evaluating the goodness of a descriptors' subset. For the evaluation, this heuristic considers both the predictive power of each descriptor individually and the level of inter-correlation between descriptors. The CFS searches for subsets that, on the one hand, contain descriptors that are highly correlated with the class and, on the other hand, are uncorrelated with each other. We have used this method for all the experiments reported here.

Discrimination model

The discrimination model used, the Sequential Minimal Optimization (SMO) is a training algorithm for support vector machines (SVM) (Platt, 1998). The basic training principle of SVMs is the construction of a hyperplane or a set of hyperplanes in a high dimensional space that separate data points into classes with maximum margins (Vapkin, 1982). SVMs look for the largest distance of the hyperplane to the nearest training data points of any class, such that the generalization error of the classifier is minimized.

Training SVM requires solving a large quadratic programming optimization problem. SMO breaks the problem down into the possible smallest programming optimization problems. These problems are solved analytically, which improves significantly its scaling and computation time. The implementation of the SMO algorithm is included in WEKA, a data mining suite with open source machine learning software written in Java (Witten & Frank, 2005).

A validation process was carried out in order to go further than the performance of the discrimination model on the available data, and to evaluate its generalization capabilities i.e., its performance when classifying previously unseen instances. To evaluate the predictive performance of the discrimination model based, the 10-fold cross-validation method was performed. In this method, the data set is randomly divided into 10 subsets or folds. Then, 9 of the folds are used for training and one for testing. This process is repeated 10

times and the final result is averaged over the 10 runs. The classification accuracy of the discrimination model is assessed by examining the F-measure[1], a weighted average of precision and recall which varies between 1 for its best value and 0 for its worst.

5.3 Experiments

In this section, we describe the machine learning experiments conducted to investigate if infant-directed speech from Brazilian and European Portuguese can be discriminated and which are the best features to achieve this; also if infant-directed singing from Brazilian and European Portuguese can be discriminated, and which are the type of features that discriminate these two.

In addition, we will verify if the type of features (rhythmic and melodic) that perform best when discriminating infant-directed speech and singing are shared by both discrimination models.

Finally, we will explore if these features are useful for another discrimination condition, an interaction context classification task, or if they are specific to the discrimination of Portuguese variants. The descriptors computed previously will be used as input to the discrimination models.

5.3.1 Discriminating between Brazilian and European Portuguese infant-directed speech

In the present classification experiment, we aim to discriminate Brazilian Portuguese from European Portuguese utterances, exploring which features exhibit the best performance. Previous studies show that European Portuguese and Brazilian Portuguese differ regarding rhythm (Frota & Vigario, 2001). Additionally, infants can distinguish between different speech rhythm classes (Nazzi & Ramus, 2003). However, these studies used adult-directed speech and not infant-directed speech.

Can these two Portuguese variants be discriminated when dealing with infant-directed speech? What are the acoustic properties that best discriminate these two Portuguese variants? Are the rhythmic distinctions between Portuguese variants still noticeable in infant-directed speech register?

We will look for acoustical correlations that can identify differences between the two Portuguese variants. Table 5.3 provides statistical information of the utterances dataset built for this experiment. Statistics reveal that the Brazilian Portuguese speech rate is higher than the European Portuguese one. This result might reflect some level of vowel reduction or even vowel suppression present in European Portuguese, given that speech rate is the measure of vocalic intervals per second.

[1] $Fmeasure = \frac{(2 \times recall \times precision)}{(recall + precision)}$

Table 5.3: Basic statistical information about the utterances grouped by Portuguese speech variant.

	Brazilian Portuguese	**European Portuguese**
Number of instances	457	464
Duration (s) Mean (std)	1.58 (0.62)	1.84 (0.74)
Speech rate (V/s) Mean (std)	3.84 (1.08)	2.99 (0.98)
Mean F_0 (Hz) Mean (std)	275.77 (74.88)	285.20 (76.13)

Attribute selection was performed with CFS in order to identify a group of relevant descriptors for the discrimination task. The selected group of descriptors is mainly composed by rhythm-related features:

- rPVI of the consonant interval durations
- Standard deviation of the vocalic interval durations
- Coefficient of variability of the consonant interval durations
- Speech rate
- Percentage of vocalic intervals with falling pitch

Table 5.4 presents the mean, standard deviation and p-value for rhythm-related descriptors shown to be relevant in language discrimination tasks (see sub-section Durational units considered), as well as pitch-related descriptors associated with the contour shape with statistical relevance. P-values were obtained performing a t-test for independent samples, with Portuguese variant as a factor and the descriptors as dependent variables.

Rhythm-related descriptors show higher statistical significance regarding the discrimination of Portuguese variants when compared with contour shape related descriptors, such as initial slope and variance of the approximation of the pitch contour. European Portuguese exhibits higher durational contrast than the Brazilian variant for the vocalic and consonant duration intervals. V-to-V durations did not show statistical relevance for discriminating between Portuguese variants.

To conclude, we ran the classification method using the sequential minimal optimization algorithm for training a support vector classifier with a 10-fold cross-validation test mode. Results achieved with the stratified 10-fold cross-validation test gave 68.3% correctly classified instances (627 correct over 291 incorrect) with an accuracy F-measure of 0.68.

Table 5.4: Mean, standard deviation and p-value for a group of features, considering Brazilian and European Portuguese speech variants.

	Brazilian Portuguese Mean (std)	**European Portuguese** Mean (std)	***p***
nPVI (V durations)	59.60 (32.71)	67.46 (37.67)	0.003
nPVI (V-to-V durations)	43.38 (28.79)	43.09 (29.66)	0.52
rPVI (C durations)	11.62 (9.40)	18.86 (16.47)	$<$0.001
CV (C durations)	0.61 (0.256)	0.74 (0.30)	$<$0.001
Initial slope of pitch contour	29.98 (418.34)	-46.79 (424.30)	0.019
Variance of the pitch contour	0.12 (0.16)	0.18 (0.28)	0.003

5.3.2 Discriminating between Brazilian and European Portuguese infant-directed singing

In this experiment, the aim is to discriminate between infant-directed singing from the Brazilian and European Portuguese samples. It is known that infants in a pre-verbal stage focus on prosodic cues present in music and speech, and may perceive these stimuli as sound sequences that follow patterns of rhythm, stress, and melodic contours (Trainor et al., 2004; Mehler et al., 1996; Nazzi & Ramus, 2003). Therefore, infants may treat both music and speech using the same perceptual processes.

Can infant-directed singing from the two Portuguese variants be discriminated using the same cues as for infant-directed speech? For the implementation of this experiment, we have followed the same steps as before so that results are comparable. We have computed the same durational units using the method described earlier and extracted the same descriptors (see sub-section Discrimination system model). Statistical information of the utterances in dataset built for this experiment is provided in Table 5.5. Once again, speech rate is higher for Brazilian Portuguese, as had occurred with for speech.

As before, we performed a CFS based attribute selection in order to identify a group of relevant descriptors for the discrimination task. The group of features shows, as in the previous experiment with speech, a strong presence of rhythm-related features:

Table 5.5: Basic statistical information about the utterances grouped by Portuguese singing variant.

	Brazilian Portuguese	**European Portuguese**
Number of instances	28	28
Duration (s) Mean (std)	7.22 (4.27)	11.99 (7.85)
Speech rate (V/s) Mean (std)	3.10 (0.63)	1.99 (0.37)
Mean F_0 (Hz) Mean (std)	263.30 (33.96)	275.75 (48.97)

- rPVI of consonant interval durations
- Standard deviation of vocalic interval durations
- Speech rate
- Percentage of vocalic intervals in which pitch rises
- Percentage of vocalic intervals in which pitch is flat
- Intermediate slope of pitch contour approximation

It can be observed that three features (rPVI of the consonant interval durations, standard deviation of the vocalic interval durations, and speech rate) are common in the selected sets of speech and singing. Table 5.6 presents the mean, standard deviation and p-values for rhythmic contrast descriptors reported in the previous experiment, as well as rhythm and pitch-related features that showed statistical significance for the discrimination of Portuguese singing variants. These results were obtained performing t-tests for independent samples, with Portuguese variant as a factor and the descriptors as dependent variables.

As observed in the speech materials, European Portuguese singing exhibits higher durational contrast than Brazilian Portuguese for the vocalic and consonantal interval durations. V-to-V durations, once again, did not show statistical relevance for discriminating the Portuguese variants.

Finally, we ran a 10-fold cross-validation experiment using the SMO classification algorithm. Results yielded 83.9% correctly classified instances (47 correct over 9 incorrect) with an accuracy F-measure of 0.83.

An additional analysis was carried out in order to assess the performance of the classification model built for speech (see Discriminating between

Brazilian and European Portuguese infant-directed speech) applied now to the singing materials.

The results for this analysis with the stratified 10-fold cross-validation test gave 67.86% correctly classified instances (38 correct over 18 incorrect) with an accuracy F-measure of 0.64. Performing the inverse analysis, that is, applying the singing model to 10 different subsets of speech materials, each one containing the double of total singing instances (2x56 = 112), we obtained 76.4% correctly classified speech instances (F-measure = 0.7601; std = 0.0393).

Table 5.6: Mean, standard deviation and p-value for a group of features, considering Brazilian and European Portuguese singing classes

	Brazilian Portuguese Mean (std)	**European Portuguese** Mean (std)	***p***
nPVI (V durations)	52.40 (12.13)	60.87 (19.37)	0.065
nPVI (V-to-V durations)	49.33 (17.88)	46.07 (14.59)	0.476
rPVI (C durations)	16.35 (10.33)	26.21 (10.99)	0.002
Std (V durations)	0.08 (0.03)	0.15 (0.05)	<0.001
%V which pitch rises	0.02 (0.03)	0.11 (0.09)	<0.001
%V which pitch is flat	0.91 (0.11)	0.7193 (0.17)	<0.001
Intermediate slope of the pitch contour	-4.72 (50.06)	20.87 (28.98)	0.03

5.3.3 Discriminating interaction classes: Affection vs. disapproval vs. questions

Previous research has shown that the shape of the melodic contours of infant-directed speech can be categorized into contour prototypes according to communicative intent (Fernald, 1989). Automatic characterization of emotional content in motherese has been implemented and features concerning the melodic contour of speech have shown satisfactory results (Mahdhaoui et al., 2009).

Do melodic contour related features show the best performance when discriminating interaction classes such as affection, disapproval and questioning? Can these interaction classes be discriminated using descriptors related with the shape of the speech melodic contour, in contrast with the discrimination of speech variants, in which rhythm-related features yielded better performance?

In this experiment, we aimed to detect the best features for the discrimination of interaction types, examining if the features used to discriminate speech and singing are specific to the discrimination of Portuguese variants, or if they are also discriminative in different conditions, namely an interaction context discrimination task.

For this experiment we have considered the three interaction contexts of affection, disapproval and questioning in a cross-Portuguese variant approach. In other words, we have grouped all the interaction units belonging to a specific interaction context, regardless of the Portuguese variant to which they pertained.

The dataset for this experiment was organized as shown in Table 5.7, that also shows the statistical information about the utterances in each class. The affection class gets the highest mean fundamental frequency value, whereas the disapproval class gets the lowest. Regarding speech rate, the question class has the highest value, and affection class the lowest.

Table 5.7: Basic statistical information about the utterances grouped by interaction classes.

	Affection	**Disapproval**	**Question**
Number of instances	313	300	308
Duration (s) Mean (std)	2.07 (0.77)	1.56 (0.61)	1.49 (0.50)
Speech rate (V/s) Mean (std)	2.91 (0.91)	3.47 (1.13)	3.85 (1.08)
Mean F_0 (Hz) Mean (std)	300.37 (79.29)	256.41 (74.02)	283.84 (66.55)

Attribute selection was performed in order to identify a group of relevant descriptors for the discrimination task. Only two features are not related with pitch and contour shape. The group of selected features includes:

- Initial slope of the pitch contour approximation
- Intermediate slope of the pitch contour approximation
- Final slope of the pitch contour approximation
- Skewness of the pitch contour approximation

- Variance of the pitch contour approximation
- Mean pitch for each utterance
- The percentage of vocalic intervals in which pitch falls
- Standard deviation of the duration of vocalic intervals
- Speech rate

One-way ANOVAs were calculated, with interaction class as factor and descriptors as dependent variables, in order to test a possible dependency of the observed descriptor values on the different communication contexts. Table 5.8 presents the mean, standard deviation and p-value for the rhythmic contrast descriptors reported in the previous experiments as well as rhythm and pitch-related features that showed statistical significance for the discrimination of the singing variants.

Finally, we have run a 10-fold cross-validation experiment as the previously reported ones. Results for this analysis yielded 63.62% correctly classified instances (584 correct over 334 incorrect) with an accuracy F-measure of 0.64.

Table 5.8: Mean, standard deviation and p-value for a group of features, considering affection, disapproval and question speech contexts.

	Affection Mean (std)	**Disapproval** Mean (std)	**Question** Mean (std)	***p***
nPVI (V durations)	70.66 (35.44)	63.37 (36.99)	56.60 (32.57)	<0.001
nPVI (V-to-V durations)	47.53 (28.42)	40.06 (31.44)	41.99 (28.04)	0.004
rPVI (C durations)	17.62 (14.82)	16.20 (15.25)	11.96 (10.60)	<0.001
Std (V durations)	0.117 (0.073)	0.068 (0.044)	0.060 (0.041)	0.001
Skewness of pitch contour	0.058 (0.328)	-0.054 (0.322)	0.112 (0.288)	<0.001
Initial slope of pitch contour	-17.29 (296.98)	79.54 (354.28)	-114.56 (468.22)	<0.001
Final slope of pitch contour	-79.49 (227.31)	-45.62 (443.04)	172.75 (451.65)	<0.001

As mentioned before, previous research has categorized communicative intents into prototypical melodic contours in infant-directed speech (Fernald,

1989). These prototypical shapes have been considered cross-linguistic universals (Papoušek & Papoušek, 1991). However, despite these cross-linguistic universals, can the different rhythmic patterns between Portuguese variants be noticeable? In other words, can the interaction classes be discriminated considering the Portuguese variant? Can the mixture of rhythmic differences between Portuguese variants and contour shape differences between interaction classes solve this discrimination problem?

We examined the predictive performance of the computed descriptors in a more complex task. In this analysis, we aim to assess the performance of the discrimination between interaction classes, but this time considering simultaneously the Portuguese variant to which each instance belongs.

We expected that the discrimination model was able to detect different interaction classes and simultaneously the Portuguese variants. Six different classes were considered: Brazilian Portuguese (BP) Affection, Disapproval and Question (A-BP, D-BP, Q-BP, respectively, in Table 5.9), and European Portuguese (EP) Affection, Disapproval, Question (A-EP, D-EP, Q-EP, respectively, in Table 5.9).

The distribution of instances per classes as well as the corresponding statistical information is shown in Table 5.9. For both the Brazilian and the European variants, the Question class shows the highest value for speech rate, as also happened in the preceding experiments. Overall results for speech rate are higher for the Brazilian Portuguese variant, when comparing equivalent interaction classes.

Table 5.9: Basic statistical information about the speech utterances grouped by classes considering interaction contexts and Portuguese variants (see text).

	A - BP	**D - BP**	**Q - BP**	**A - EP**	**D - EP**	**Q - EP**
Number of instances	151	150	156	162	150	152
Duration (s) Mean (std)	2.01 (0.63)	1.36 (0.52)	1.39 (0.46)	2.13 (0.89)	1.76 (0.63)	1.59 (0.52)
Speech rate (V/s) Mean (std)	3.34 (0.85)	3.88 (1.14)	4.27 (1.02)	2.51 (0.78)	3.07 (0.98)	3.41 (0.96)
Mean F_0 (Hz) Mean (std)	284.64 (77.22)	258.60 (81.44)	283.69 (62.65)	315.03 (78.62)	254.21 (65.98)	283.99 (70.53)

In the additional analysis, we performed an attribute selection in order to identify a group of relevant descriptors for the discrimination task. The presence of rhythm-related features is stronger for this discrimination problem

as compared to the set of features selected in the previous one:

- Initial slope of the pitch contour approximation
- Intermediate slope of the pitch contour approximation
- Final slope of the pitch contour approximation
- Variance of the pitch contour approximation
- The percentage of vocalic intervals in which pitch falls
- Mean pitch for each utterance
- rPVI of the consonant interval durations
- Standard deviation of the vocalic interval durations
- Speech time
- Speech rate

We ran several ANOVAs to test the effect of language variant and interaction context (and their possible interaction) on each descriptor listed above, and found that in most of the cases only the effect of the interaction context was statistically significant ($p < 0.001$). This was observed for 7 descriptors (5 pitch-related and 2 rhythm-related), namely initial slope of the pitch contour approximation ($F = 20.42; d.f. = 2$), intermediate slope of the pitch contour approximation ($F = 4.80; d.f. = 2$), final slope of the pitch contour approximation ($F = 38.42; d.f. = 2$), variance of the pitch contour approximation ($F = 42.64; d.f. = 2$), mean pitch for each utterance ($F = 28.48; d.f. = 2$), std of vocalic intervals duration ($F = 97.36; d.f. = 2$) and speech time ($F = 92.23; d.f. = 2$).

For 3 descriptors (1 pitch-related and 2 rhythm-related) only the variant was significant, namely vocalic intervals in which pitch falls ($F = 47.18; d.f. = 1$), rPVI of the consonant interval durations ($F = 66.96; d.f. = 1$) and speech rate ($F = 166.74; d.f. = 1$).

Finally, we ran a 10-fold cross-validation experiment analogous to the previous ones. Results for this analysis yielded 46.73% correctly classified instances (429 correct over 489 incorrect) with an accuracy F-measure of 0.46. As can be seen in Table 5.10, that shows the confusion matrix, communicative contexts are confused across variants.

Table 5.10: Confusion matrix for the classification considering interaction speech contexts and Portuguese variants.

A - EP	D - EP	Q - EP	A – BP	D - BP	Q - BP	Classified as
104	17	10	23	2	6	A – EP
24	**55**	10	21	26	13	D – EP
21	18	**54**	12	17	30	Q – EP
27	26	13	**69**	7	9	A – BP
4	30	13	10	**58**	33	D – BP
2	9	17	11	28	**89**	Q – BP

5.4 Discussion

The present study explored rhythmic and melodic patterning in speech and singing directed to infants from Brazilian and European Portuguese variants. Different classification configurations were conducted in order to provide insight into the prosodic characterization of the infant-directed register of speech and singing from the two Portuguese variants.

In the first experiment, Brazilian and European Portuguese infant-directed speech were discriminated with a 68.3% success rate. The attribute selection performed identified a group of the five best features in which four were rhythm-related, demonstrating strong predictive power. The results indicate that there are relevant rhythm differences between infant-directed speech from the two Portuguese variants and not melodic differences; durational contrasts are higher in European Portuguese than in Brazilian Portuguese (see nPVI and rPVI values in Table 5.4).

As referred before, the two Portuguese variants are considered to have distinct rhythm patterning (Frota & Vigario, 2001): European Portuguese is considered more stress-timed, characterized by vowel reduction and, therefore, with higher durational contrast values and, contrastingly, Brazilian Portuguese is considered more syllable-timed. Therefore, despite a natural tendency in infant-directed speech to clearly articulate phonemes, namely vowels, in order to facilitate language acquisition (Papoušek et al., 1987), a different rhythm patterning is still observable between the Portuguese variants. These results demonstrate that both variants keep rhythm patterning differences in the infant-directed speech register.

It would be of interest to test the same discriminative features found in this experiment for discrimination between adult-directed speech from the same two Portuguese variants. Should the same features not reveal the same discriminative power for adult directed speech, it would be important to determine if these features are "infant-adapted" and to explore adaptive explanations for this fact.

In the second experiment, Brazilian and European Portuguese infant-

directed singing were discriminated with 83.9% success rate. The set of features identified by an attribute selection includes six features, in which half were rhythm-related and half were pitch-related. The three rhythm-related features, namely rPVI of consonant interval durations, standard deviation of the vocalic interval durations and speech rate, were also part of the group of features with high predictive performance built for the speech materials.

Moreover, the model trained with speech is capable of correctly classifying 67.86% of the singing materials, and the inverse analysis applying the singing model to speech materials yields 76.4% correctly classified instances. These results considering the discrimination between language variants indicate that processing speech and singing share the analysis of the same properties of the stimuli.

Additionally, values for durational contrasts in singing are higher for the European Portuguese materials (see nPVI and rPVI values in Table 5.6), as observed with infant-directed speech. Therefore, rhythmic patterning differences are also kept in the singing material. These results are consistent with previous findings relating the musical rhythm of a particular culture with the speech rhythm of that culture's language (Hannon, 2009; Patel et al., 2006)).

Our last experiment examined the discrimination between pragmatic classes such as affection, disapproval and questioning, and the resulting model correctly classified 63.6% instances. In this experiment, pitch-related features revealed to be efficient for the pragmatic discrimination, in contrast to what had been observed for the language variant discrimination.

When we look at the simultaneous detection of interaction and variant, the presence of rhythm-related features as the best descriptors for the task is noticeable. This contrasts with the set of features required for the discrimination between variants only, or between interactions only, where few rhythm descriptors were needed.

A closer analysis of the confusion matrix produced by this classification problem reveals that the communicative contexts were similar across variants and therefore they yielded many classification confusions. This confirms the presence of cross-linguistic properties of different interaction contexts (Papoušek & Papoušek, 1991).

Summing the correctly classified cases in each interaction context irrespective of language variant (for example, the 104 correct cases from European Portuguese affection plus the 23 cases from Brazilian Portuguese affection, and so on, cf. Table 5.10), would make a total of 582 cases. Therefore, disregarding errors in classifying language variants, we get a 63.4% successful discrimination of interaction contexts, a value closer to the one obtained in the classification problem where only the interaction classes were considered.

Another fact worth being noted is that the speech rate values, for all the experimental set-ups, are found to be higher for the Brazilian Portuguese variant. Speech rate was measured here as the number of vocalic intervals

per second. Therefore, this result might reflect some level of vowel reduction or even vowel suppression in European Portuguese, which could in turn imply that certain vocalic intervals are absent in this variant.

Additionally, vocalic and consonantal intervals revealed to be more relevant in comparison to the V-to-V durations for discriminating the Portuguese variants. These results are consistent with previous findings suggesting a rhythm based discrimination by newborns relying on distinctions between vowels and consonants (Mehler et al., 1996; Nazzi & Ramus, 2003; Ramus et al., 1999).

Although the main goal of this study was not focused on the robustness of the discrimination models, but rather on the results of these models as a means to capture rhythmic and melodic patterns in speech and singing directed to infants, the classification results for all experimental configurations were below our expectations. It is possible that, for an automatic discrimination approach such as the one adopted here, more instances were needed or that the materials do not contain the equivalent text (sentences) for each variant. It could also be the case that the features used were not sufficiently efficient. An effort should be made in the future in the sense of exploring more descriptors for the discrimination tasks performed in this study.

Finally, care has been taken in collecting representative stimuli of what is most salient to an infant, that is, infant-directed speech and singing, and descriptors have been computed trying to capture the perception and processing of prosodic patterns from the perspective of an infant. Therefore, the results achieved may reveal that prosody of the surrounding stimuli of an infant, such as speech and singing, is a source of rich information not only to make a distinction between different communicative contexts but also to provide specific cues about the prosodic identity of their mother tongue.

5.5 Conclusion

The main goal of the present study was to explore rhythmic and melodic patterning in speech and singing directed to infants from Brazilian and European Portuguese variants. Different machine learning experiments were conducted in order to provide insight into the prosodic characterization of the infant-directed register of speech and singing from the two Portuguese variants.

Descriptors related with rhythm, namely rPVI of the consonant interval durations, standard deviation of the vocalic interval durations and speech rate, showed strong predictive ability for the discrimination of the Portuguese variants, both in speech and in singing. Moreover, different rhythmic patterns were observed in the two variants, with higher durational contrasts for European Portuguese speech and singing than for Brazilian Portuguese (see nPVI and rPVI values in Table 5.4). Further investigation should be carried out to determine if these prosodic differences are related to infant

development of musical predispositions and how they might bias melodic representations differently for each culture.

Rhythm-related descriptors were not relevant for the discriminations of interaction contexts. However, when increasing the complexity of the interaction classification problem by including the language variants, rhythm-related features emerged as more relevant than they had been in the context-only classification problem. Therefore, we provide additional evidence that prosody of the surrounding stimuli of an infant, such as speech and singing, are rich sources of information to make a distinction between communicative contexts through melodic information, and also to provide specific cues about the rhythmic identity of the native language.

Moreover, common features were used by the classification method for discriminating speech and singing tasks. This indicates that processing speech and singing share the analysis of the same properties of the stimuli. Hence, these results strengthen previous findings by providing further evidence that the cognition of music and language may share computational resources during the pre-verbal period.

We consider that, rather than recognizing or discriminating, such as the approach taken in this study, the infant has to learn patterns and discover structures. Consequently, in the next research step we will build a developmental model for exploring the fact that prosodic features present in infant-directed speech and singing may affect the infant's development of rhythmic representations.

6

Temporal information processing in early development: a computational model for exploring correlations and possible interactions between music and speech

Summary

In this chapter we describe a computational model based on temporal difference reinforcement learning used to explore how the temporal prosodic patterns of a specific culture influence the development of rhythmic representations and predispositions in human infants. With this purpose, we performed the computational simulation of Yoshida et al. (2010)' s empirical experiments.

The model is composed by two main modules: (1) the perception module, where the raw auditory data is processed in a way analogous to early sound processing, and (2) the representation module consisting of the neural network that learns to classify sounds on the basis of the output of the perception module. The training data is composed by three categories of sounds: (1) infant-directed speech, (2) infant-directed singing, and (3) environmental sounds.

We assessed the extent to which the model's exposure to the durational patterns of a specific culture influenced the construction of its internal representations and if this construction changed with greater auditory experience.

Across development, the model showed a similar developmental profile as that found in infants. Moreover, infant directed singing was found to be a crucial factor in explaining this developmental profile.

We conclude that the exposure to the surrounding sound environment influences the development of auditory temporal representations and that the singing may work as a facilitator in learning the temporal regularities of a specific language and auditory culture.

6.1 Introduction

Musical capacity is universal to all cultures. Despite its diversity, it is present in every human society and in every historical period (Nettl, 2000). However, there are important cultural differences regarding the processing and representation of music (Morrison et al., 2003). These differences might be due to many culturally specific factors, especially those related with the early exposure to the melodic, harmonic and rhythmic features that compose the auditory environment of a specific culture. This early exposure may lead to the development of particular cultural perceptual constraints (Morrison & Demorest, 2009).

Among the variety of sonic experiences that can influence human auditory perception, music and speech have been highlighted as determinant factors (Carterette & Kendall, 1999; Patel et al., 2006; Iversen et al., 2008; Hannon, 2009). Of the auditory information to which infants are exposed, the most salient are speech and singing sounds (Nakata & Trehub, 2004; Cooper & Aslin, 1994; Trehub et al., 1993). Moreover, from the perspective of a pre-verbal infant, music and speech may both be perceived as sound sequences that unfold in time, following patterns of rhythm, stress, and melodic contours (Masataka, 2009). Therefore, despite the surface differences of speech and musical stimuli, infants may nevertheless use the same cognitive processes to attend to the melodic and rhythmic aspects of both music and speech (Patel, 2007).

Music and language cognition, and their interactions have been studied using different approaches and techniques. For example, Patel et al. (2006) studied language and music relations by quantitatively comparing rhythms and melodies of speech and instrumental music. This investigation has shown that instrumental music (rhythms and melodies) reflects the prosody of a composer's native language. Also supporting the suggestion that musical rhythm of a particular culture may be related with the speech rhythm of that culture's language, Hannon (2009) showed that subjects can discriminate between instrumental songs composed in two languages that have different rhythmic prosody basing their decisions on rhythmic features only.

Another example of the interaction of these two domains comes from rhythm perception, specifically from auditory perceptual grouping, a fundamental operation in processing temporal elements from auditory patterns. Based on the Gestalt notion of the integration of parts to form a whole, perceptual grouping consists in the process of selecting components from an analysed scene that, by being grouped, form an individual object. Similarly, the mechanisms of organization, segmentation and grouping occur in the auditory domain as well, involving spatial and temporal segregation of events (Bregman, 1994). For the matter, we are interested in the temporal, specifically, rhythmic grouping. Grouping mechanisms are subjacent in the perception of rhythm. These mechanisms are motivated by the detection of stressed or salient events, in order to segment into coherent higher-level

patterns (e.g., motives and phrases). The salience or prominence in a stream of events is perceived through fluctuations in loudness, duration and pitch values of the rhythmic elements. Either in speech or in music, the rhythmic element is stressed by the joint occurrence of (i) peaks of loudness together with long duration intervals or (ii) peaks of loudness together with peaks of pitch (Bolton, 1894; Woodrow, 1909). Combinations of stressed events, non-stressed events, and silences create the sensation of boundaries. These boundaries lead to the perceptual segmentation of rhythmic events, also referred as rhythmic grouping.

Auditory perceptual grouping is a central component in rhythmic behaviours such as music and language. These two structures are both temporally organized and unfold over time. Therefore, its processing depends on abilities to process temporal information. The grouping of these two structures will determine how we segment a continuous stream of sound into smaller pieces, building a complex interpretation of the acoustic input.

Perceptual grouping processes are known to be operative early in infancy (Thorpe & Trehub, 1989), and the principles governing these processes have been proposed to be universal, forming an innate building block of perception. These principles, similar to what underlies the perception of visual patterns and its link to the gestalt principles, follow two main rules: (i) louder sounds tend to mark group beginnings and; (ii) longer sounds tend to mark group endings. These rules are known as the Iambic and Trochaic laws respectively (Bolton, 1894; Woodrow, 1909). In addition, these principles are considered as key causal factors in language learning, since they influence aspects of language learning such as word segmentation.

However, recent studies have shown that perceptual grouping is not universal but learned from the environment and thus dependent on experience (Iversen et al., 2008). Also in contrast with much previous research, Yoshida et al. (2010) demonstrated a cultural difference in non-linguistic perception of grouping. This difference develops early due to the different language environments. In two experiments, they tested Japanese and English learning infants of 5-6 and 7-8 months. They observed that 5-6 month old infants do not reveal any systematic perceptual grouping bias (e.g., preference for grouping based on English or on Japanese patterns). In contrast, 7-8 month old infants developed grouping preferences. These preferences were consistent with the ones found in adulthood (Iversen et al., 2008). The key factor for explaining the development of different grouping preferences is that these infants (American vs Japanese) have different auditory experiences. Additionally, the most salient cause of cultural differences in auditory experience is the dominant language, which also explains that the grouping preferences that are developed are consistent with their language structure and therefore reflect the rhythms of the two languages (English vs. Japanese). The learning of the duration patterns is made through the exposure to the rhythmic units present in speech. This exposure leads to an implicit learning of the rhythmic structure of speech and is responsible for shaping low-level group-

ing biases.

Other than speech, infants are also exposed to a special selection of music, consisting of lullabies and play songs (Trehub & Trainor, 1998). Parents and caregivers, across cultures, languages and musical systems, use singing to regulate infant states and to communicate emotional information (Trehub et al., 1993; Rock et al., 1999). Therefore, it is possible that speech, together with singing, is shaping the low-level grouping biases observed in the older infants. But how does the structure of these two sound systems, namely, speech and singing, influence the construction of rhythmic preferences and representations? And how do they interact in this process? How important is it that singing is present, in addition to speech?

In this context, aiming to explore these questions and, at the same time, produce explanations for how the temporal prosodic patterns of a specific culture influence the development of rhythmic representations and predispositions, we propose to build a computational model based on temporal difference information processing and representation. The model is exposed to an audio training data set, typical of that which infants experience, consisting of recordings of caregivers interacting with babies, infant-directed speech and infant-directed singing utterances from European Portuguese as well as recordings of environmental sounds, all of which can be categorized by capturing temporal patterns present over durational sequences. Model performance is then evaluated against that of human infants.

For the validation of the model, we propose the simulation of an empirical experiment from Yoshida et al. (2010) that should produce similar results. Furthermore, on Experiment 1 (see section 6.4.2), we aim to verify if the model's exposure to the durational patterns of a specific culture influence the construction of its internal representations and if this construction is developmental dependent.

In Experiment 2 (see section 6.4.3), we perform manipulations in the input environment, with the aim of deriving predictions from the model. Thereby, we exclude the singing material with the aim of testing its influence in the process of the model's building of rhythmic internal representations and preferences. After a brief overview of the model, the experiments are reported in section 6.4. In section 6.5, we present a general discussion of the experiments, and, finally, the last section (6.6) presents our conclusions.

6.2 Model overview

The model is composed of two main modules, the *perception* and the *representation* modules (see Figure 6.1). In the perception module, the raw data is processed for input to the representation module. Operations such as durational intervals extraction from the audio data and subsequent pitch, duration and loudness are computed. The representation module consists of a neural network and it is fed with the data that was pre-processed. Each

of these is discussed in more detail below.

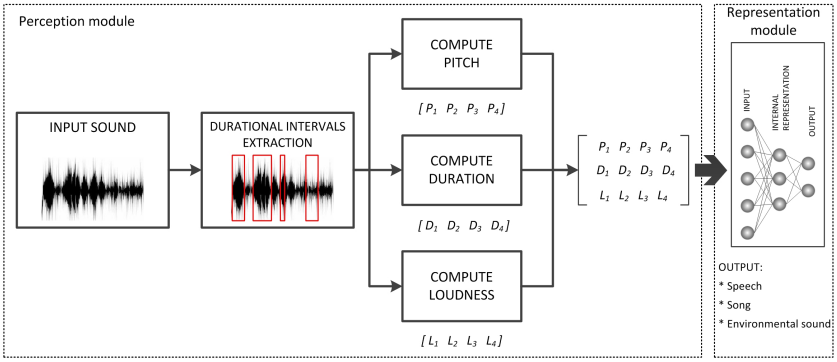

Figure 6.1: Schematic illustration of the computational model.

6.2.1 Perception module

Infants are very sensitive to rhythmic information. During the pre-linguistic period, when developmental processes are particularly susceptible to environmental input, rhythmic cues from prosody play an important role in acquisition, either in music or speech, as they provide information to segment continuous streams into meaningful units and learn about their structures.

Vowels are perceptually relevant regarding rhythm. They are especially important in languages with rhythmic patterns close to stressed-timing, where stress has a strong influence on vowel duration and the marking of certain syllables within a word as more prominent than others leads to vowels' duration fluctuation. Infants are able to segment vocalic intervals from the speech stream (Ramus et al., 1999). Ramus et al. (1999) argue that infants perform a crude segmentation of the speech stream which only distinguishes vocalic and non-vocalic portions, and classify different languages based on this contrast. Vocalic intervals are acoustically characterized by the portions of the signal that exhibit periodic spectral patterns over time. This way, for the model, we consider a durational interval as a segment of signal that shows constant spectral behaviour over time, whether it is the case of speech or singing audio materials or not. Therefore, the same method for extracting the durational intervals is used for all the audio materials.

Both in speech or music, the fluctuations in pitch, duration and loudness values are perceived as salient or prominent events in a stream. Combinations of stressed events, non-stressed events and silences create the sensation of boundaries. These boundaries lead to the perceptual segmentation of the events, also referred as rhythmic grouping (Palmer & Hutchins, 2006). This way, we computed the significant components that characterize prosody,

from each durational interval, that is, pitch duration and loudness (Nooteboom, 1997).

The same processing methods, including the extraction of the durational intervals, are used for all the input data, regardless of its category (i.e., speech, singing or environmental sounds). Indeed, for infants without any verbal knowledge, music and speech may not be as differentiated as for older children and adults and therefore it is possible that they are using the same cognitive operations and sharing processing resources (Patel, 2007).

The result of this processing, the module outputs a vector composed by the durational events of each utterance on the audio dataset. Each durational event is composed by its duration, pitch and loudness values (see Figure 6.2 on Data Section).

6.2.2 Representation module

The model is situated within connectionist framework. Connectionism, from a methodological perspective, can be seen as a structure for studying and explaining cognitive phenomena by means of artificial neural network models. Neural network modelling is an excellent tool for understanding development, for exploring the link between multiple interacting biological and environmental constraints, and the development of cognitive representations (Shultz, 2003; O'Reilly & Munataka, 2000; Haykin, 2009). Based on the principle that human brain is a complex, nonlinear and parallel information processing system, neural network models are designed under a neurobiological analogy with the brain, composed by neuron-like simple processing units. These units are functionally related by synaptic-weight, and adapt to the input data, capturing the plasticity of a developing nervous system to the surrounding environment (Haykin, 2009). In these models, processing and memory are distributed over a whole network, exploiting the parallel processing capability of the human brain. This processing consists on the propagation of activation through the network of simple processing units that are interconnected by weighted connections. This way, the internal representation of knowledge can be observed in the activation values of the hidden processing units. Consequently, knowledge is represented in the structure of the processing system, in opposition to symbolic approaches, where knowledge is transferred between different memory registers. Modifications in the connections, driven by experience, provide a mechanism for both learning and development. This way, gradual learning and development can take place as small changes that shape the values of the activations and the weighted connection. As a product of this learning process, connectionism proposes that cognition emerges in neural network models (Hinton, 1989). This approach has been successfully applied to modelling many developmental phenomena (Mareschal & Thomas, 2007).

Therefore, connectionist models have been a powerful method for computing learning and development, motivated by plausible biological support.

These models provide the possibility of implementing graded representations, where response can assume any of a range of continuous values as it happens in human knowledge (Mareschal et al., 2007).

The model uses the Temporal Differences (TD) learning rule (Sutton, 1988). This method belongs to a class of incremental learning procedures that uses past experience to predict its future behaviour. Instead of being driven by the error between predicted and actual outcomes, TD uses the difference between temporally successive predictions and, thus, learning happens whenever there is a change in prediction over time. The pairwise approaches that include supervised learning methods with the typical "input-output" behaviours and also prediction problems where the pair can be seen in the data based on which a prediction must be made versus item to be the actual outcome, ignore the sequential structure of the problem. In contrast, TD learning considers temporal sequences of observations and predictions, where learning proceeds simultaneously with processing, being suited for problems which data is generated over time. In addition, TD learning is combined with reinforcement learning (O'Reilly & Munataka, 2000). This algorithm provides a strong fit to biological properties in the sense that considers that an organism can produce actions in an environment and this environment, in result of these actions, produces rewards, most often delayed. As a consequence, the organism will look to produce actions that result in the maximum total amount of reward. The value function expresses mathematically such phenomena and it is:

$$V(t) = \langle \gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) ... \rangle$$

Where $V(t)$ corresponds to the value of the current state at a given point in time. The discount factor, γ, establishes the degree of how much future rewards are ignored. This value that can vary between 0 and 1, expresses the less priority that an organism gives to rewards that are distant in the future and, therefore, the degree of discount of these later rewards is bigger. $r(t)$ is the reward at time t. This way, the algorithm will map situations to actions so as to maximize numerical reward signals. This means that the model is not being told which action to take but must discover which actions yield the most reward by trying them. This is a different scenario from that of supervised learning, where learning is made from examples provided by a knowledgeable external supervisor. These actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards, leading the model to be able to learn from its own experience.

In the simulations reported below, the task involves capturing patterns over a durational sequence. Thus, the model must have the ability to consider short-long sequences. The properties described above make this model suited for the experimental scenario and developmental learning processes we aim to simulate. On one hand, reinforcement learning enables the simulation of the implicit learning that we want to achieve (i.e., made through

the exposure to the auditory context) and, on the other hand, allows solving problems that involve learning the detailed timing of events, and not just their order, (i.e., temporal contingencies that span many time steps).

6.3 Data

The training data set was constructed by considering what are the most salient sounds among all auditory information to which infants are exposed. Specifically, it contains infant-directed speech and infant-directed singing from European Portuguese. These distinctive registers are used by parents and caregivers, across cultures, languages and musical systems (Papoušek & Papoušek, 1991; Trehub et al., 1993). This audio collection is part of a database described in detail elsewhere (Salselas & Herrera, 2010). All audio signals were digital, stereo, 16 bit sampled at 44100 Hz. The recordings contain caregivers interacting with their healthy babies aged up to 18 months. During the recordings, caregivers were interacting with the babies in their home and in different contexts such as playing, feeding, bathing and putting the babies to bed. The materials contain spontaneous interactive speech and singing. The audio from the recordings were subsequently cut into utterances that we refer to as interaction units. Each of these utterances represents an instance in the audio database.

A third category of sounds was also collected. Here, *environmental sounds* belonging to the auditory environment of an infant that are distinct from speech and singing, such as animals (dog barking, bird singing), house noises (door opening, keys, tv, etc...) and outdoors noises (nature, traffic, etc) were gathered. Each of these sounds represents an instance in the audio database. The purpose of this category is, in an abstract way, to complement the collection of sounds to which the model is exposed to, with elements that are contrastive with infant-directed speech and singing, allowing the model to learn from their structure. In this way, the auditory environment of the model is enriched.

The complete database consisted of 197 instances:

- 94 Portuguese infant-directed speech instances
- 45 Portuguese infant-directed singing instances
- 58 Portuguese environmental sounds instances

Because infant-directed speech was recorded in the context of spontaneous interactions, it was very difficult to select portions of audio that belonged to a given interaction class and that were not mixed with background noise such as babbling and noise from the baby's toys. Consequently, the amount of data (number of instances) is somehow limited. However, the data is truly spontaneous and was collected from recordings from four different interaction contexts. The encoding of each audio input was organized as

shown in Figure 6.2, where D stands for the values of the durational intervals that were extracted (see Section 6.2.1), P is the median extracted pitch and L the median extracted loudness of the respective durational interval.

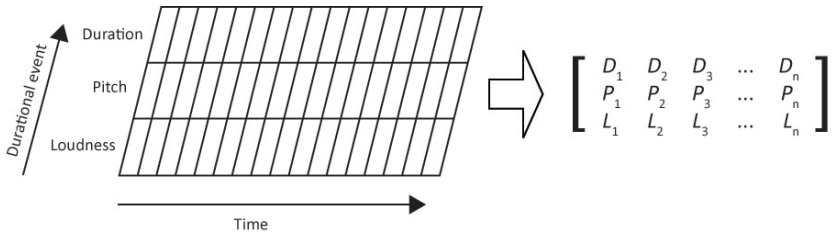

Figure 6.2: Representation of the input data. D stands for duration, P stands for pitch and L stands for loudness. The azimuthal dimension represents each durational interval of an instance made up of the three values D, P & L. In the horizontal direction, values are placed by the order of temporal extraction, in accordance with their index (1 to n, where n can be up to 33).

This form of organizing the data is motivated by the assumption that duration, pitch and loudness are processed in parallel in the brain. This way, each column represents one durational interval and, additionally, the columns imprint the sequential dimension in time.

In addition to the training data, test and familiarization input stimuli are needed to evaluate the model. Throughout the simulations, the model is submitted to a familiarization phase and tested after that. These were designed to reflect the structure of the experiment described by Yoshida et al. (2010). In their experiment, tones had a fundamental frequency of 256 Hz with 67 dB. Therefore, in our encoding, tones are differentiated from the inter-stimulus intervals (pauses) by their pitch and loudness values. In tones, pitch and loudness are 256 and 67, respectively, while in inter-stimulus intervals, these values are zero. These stimuli are encoded according to the representation illustrated in Figure 6.2. Examples of these stimuli, namely, familiarization and test (iambic and trochaic) stimuli are shown in Table 6.3. In each stream of tones (horizontal direction), each tone is composed by 3 components, namely, duration, pitch and loudness (vertical direction).

6.4 Simulations

In this section, we will describe the computational experiments done on temporal information processing in music and speech. Our first step was to simulate Yoshida et al. (2010)' experiment carried out with English-learning infants.

In their study, 5- to 6-month-old and 7- to 8-month olds are familiarized with a sequence of tones of different durations (short-long sequences). After

Table 6.1: Examples of iambic, trochaic and familiarization stimuli.

$\begin{bmatrix} 0.2 & 0.2 & 0.6 & 0.4 & 0.2 & (...) \\ 256 & 0 & 256 & 0 & 256 & (...) \\ 67 & 0 & 67 & 0 & 67 & (...) \end{bmatrix}$	**Iambic (short-long sequence) stimuli representation**. In this sequence, short (0.2 s) and long (0.6) tones alternated, with inter-stimulus intervals 0.4s long after the long.
$\begin{bmatrix} 0.6 & 0.2 & 0.2 & 0.4 & 0.6 & (...) \\ 256 & 0 & 256 & 0 & 256 & (...) \\ 67 & 0 & 67 & 0 & 67 & (...) \end{bmatrix}$	**Trochaic (long-short sequence) stimuli representation**. In this sequence, short (0.2 s) and long (0.6) tones alternated, with inter-stimulus intervals 0.4s long after the short.
$\begin{bmatrix} 0.2 & 0.2 & 0.6 & 0.2 & 0.2 & (...) \\ 256 & 0 & 256 & 0 & 256 & (...) \\ 67 & 0 & 67 & 0 & 67 & (...) \end{bmatrix}$	**Familiarization stimuli representation**. In this sequence, short (0.2 s) and long (0.6) tones alternated, with all inter-stimulus intervals 0.2s long.

this initial stage familiarization, during a test phase, infants' preferences for iambic or trochaic durational sequences were assessed. This was done using a head-turn preference paradigm, where looking time is measured preceded by a familiarization phase (Kemler et al., 1995). Depending on the pre-exposure time to the familiarization stimulus, infants change their orienting preferences from familiar to novel stimuli, a behaviour that is exploited by the habituation paradigm (Houston-Price & Nakai, 2004). This way, when interpreting the results, Yoshida et al. (2010), according to the habituation and head-turn preference paradigms, they consider that infants look longer to novel stimuli and less to familiar stimuli (Hunter & Ames, 1988). Yoshida et al. (2010) found that 5- to 6-month-olds did not show any preference for any iambic or trochaic durational pattern. However, 7- to 8-month-olds developed a preference for trochees, suggesting that this was the novel sound pattern and, thus, that iambic was the familiar sound pattern for these infants. Looking time was lower for iambs than for trochees during the test trials for the 7- to 8-month-olds.

The network's performance is evaluated by analysing its internal representations. In this analysis, we look to measure the degree of familiarization of the model to each type of test stimuli, in the same way as it is done in the behavioural experiment, and not the ability of the model to distinguish or categorize the stimuli. We will therefore focus on the model's hidden layer. Specifically, we consider the hidden unit activation vector for a given in-

put pattern to be the network's internal representation of the given pattern (Plunkett & Elman, 1997). An example of an activation vector, for a $k \times j$ dimensional hidden layer for speech is given next:

$$\begin{bmatrix} sp_{00} & \dots & sp_{0j} \\ \vdots & \ddots & \vdots \\ sp_{k0} & \dots & sp_{kj} \end{bmatrix}$$

The input data will produce activations on the hidden units and inputs that are treated as similar by the network will produce internal representations that are similar, that is, internal representations that have closer Euclidean distances to each other (Plunkett & Elman, 1997). Therefore we used Euclidean distances between hidden units activation vectors as a measure of the familiarity with each of the iambic and trochaic durational patterns. Euclidean distances were calculated according to the following equation:

$$D_{sp-i} = \sqrt[2]{\left(\begin{bmatrix} sp_{00} & \dots & sp_{0j} \\ \vdots & \ddots & \vdots \\ sp_{k0} & \dots & sp_{kj} \end{bmatrix} - \begin{bmatrix} i_{00} & \dots & i_{0j} \\ \vdots & \ddots & \vdots \\ i_{k0} & \dots & i_{kj} \end{bmatrix} \right)^2}$$

where D_{sp-i} is the distance between speech (sp matrix) and iambic (i matrix) test stimuli representations.

The distances aim to quantify how close or distant the representations of the test stimuli are from those for speech, singing and environmental sounds categories. If, for example, the representation of a given test stimulus is closer to the representation of speech, we consider that the model relates this test stimulus more with the speech category, and so forth.

This way, four distances were considered, namely:

- D_{sp-i} : distance between speech and iambic test stimuli representations;
- D_{es-i} : distance between environmental sounds and iambic test stimuli representations;
- D_{sp-t} : distance between speech and trochaic test stimuli representations;
- D_{es-t} : distance between environmental sounds and trochaic test stimuli representations;

In the experiments performed, age is manipulated by the amount of exposure to sound events (McClelland & Jenkins, 1991). With the intention of observing a developmental trajectory, the model is analysed in two different stages of learning. Thus, the model is tested twice. The first test is performed after 100 epochs have elapsed and the second test is performed after 500 epochs 500 have elapsed.

6.4.1 Experimental setup

Our aim was to follow the experimental procedure used with infants as closely as possible (see Figure 6.3).

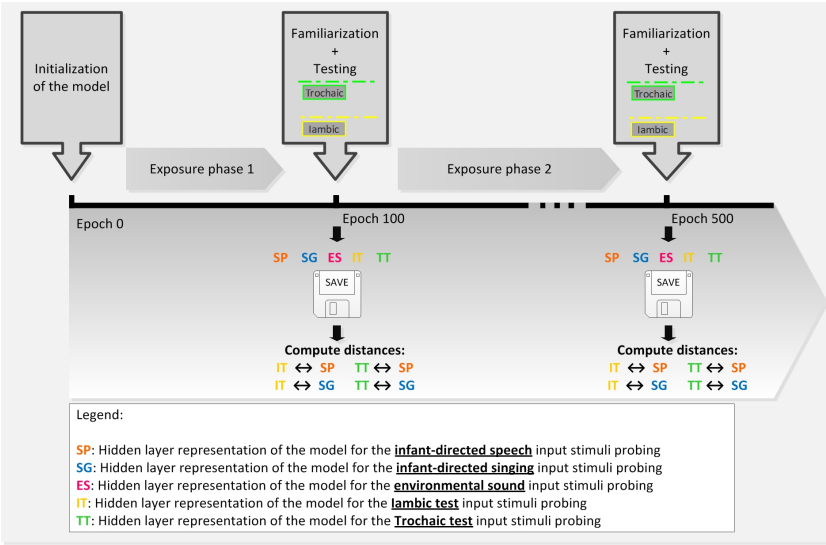

Figure 6.3: Setup used throughout the experiments.

These steps are as follows:

1. The model is initialized and a network instance is created. This was achieved by seeding the network with random initial weights and activation values, thereby simulating an individual infant to be tested. Specifically, The model is initialized and run 80 times, producing 80 different and independent results, as with independent 80 participants. These results are then averaged and statistics are computed across the 80 participants.

2. The model is exposed to the input data (infant-directed speech, infant-directed singing and environmental sounds) for a total of 500 epochs.

3. The model is tested at two different moments (100 and 500 epochs) during the exposure time. These points are meant to capture the different levels of language auditory exposure 5- and 7-month-olds will have with their general auditory environments (see Schafer & Mareschal (2001) for a similar procedure). Testing proceeded as follows:

    a) The representation across hidden layer of the model is recorded, for each category of input data, namely, infant-directed speech, infant-directed singing and environmental sounds. Specifically,

this means that a vector containing numeric activation values of the hidden layer of the model is saved at that specific moment.

b) On every occasion that the model is submitted to a test trial, a familiarization stage precedes this event, likewise in the behavioural experiment (Yoshida et al., 2010). Familiarization stage consists in exposing the model to the familiarization stimuli. The familiarization stimuli replicate the stimuli used in the original behavioural experiment concerning the characteristics of the tones (duration, fundamental frequency and loudness) and the total duration of the trials.

c) The model is exposed to the test stimuli, either iambic or trochaic that, in the same way as the familiarization stimuli, replicating the stimuli used in the original behavioural experiment.

d) After each test trial is presented to the model, for both iambic and trochaic, a representation of the hidden layer of the model is recorded.

4. Finally, the response of the model to the test stimuli to which it is exposed is analysed. In this analysis, we are interested in measuring familiarity in terms of how similar are the test stimuli with each of the input data that the model was previously exposed to in order to understand how the model is organizing the test stimuli (as infant-directed speech, singing or environmental sound). For that reason, Euclidean distances are computed between of each the representations of the test stimuli and the target speech, singing and environmental sound categories.

6.4.2 Experiment 1: Simulating a developmental trajectory in European Portuguese

RATIONALE

Music and language are the two most prominent structures in the auditory environment of infants. The exposure to these structures develops in humans non-linguistic grouping preferences consistent with their cultural temporal structure (Iversen et al., 2008). Culture-specific perceptual biases, such as sensitivity to language-specific rhythms, emerge during infancy and may be acquired by being exposed to the speech and music of a particular culture (Nazzi & Ramus, 2003). In their experiments, Yoshida et al. (2010) demonstrated that infants growing up in different language environments develop different non-linguistic grouping preferences that are consistent with their respective language's structure.

The European Portuguese language's structural properties are close to those of the English language, largely at the phrasal level. Similarities include: (i) phrasal structure (iambic: short-long); (ii) word order, that is

correlated with phrase level prosody (VO: verb-object) and; (iii) the phrasal prominence realizations (weight-sensitive stress) that is considered to be Head-Complement Right-edge (Dryer & Haspelmath, 2011). In other words, phonological prominence is stress-final, marking an iambic pattern at the phrasal level (Nespor et al., 2008)

Therefore, if English-learning infants showed higher familiarity with iambic durational sequences and this familiarity is highly influenced by the language's rhythmic structure, then a model that is exposed to European Portuguese infant-directed speech and singing should develop infant-directed speech representations that are closer (i.e., more similar) to the iambic test stimuli representations.

Thus, in this experiment, we were interested in observing the evolution of the representations that the model produces for infant-directed speech and environmental sound categories of the input data and how these representations evolve in relation to the representation produced for the iambic test stimuli.

The model was exposed to infant-directed speech, singing and environmental sounds. Through the exposure, the model captures the input data's patterns, building its internal representations. In order to assess the model's familiarity with iambic durational sequences and produce, simultaneously, results that are comparable with the empirical data, the distance between speech representations and iambic test trial representations are measured and monitored. So to establish a comparison basis for this distance and its developmental behaviour, the distance between environmental sounds' representation and the iambic test trial representation are also displayed.

RESULTS

Figure 6.4 shows the average Euclidean distance between the average hidden unit patter for iambic and environmental sounds. Figure 6.5 illustrates the same relations but between trochaic and environmental test trial representations.

Model performance was tested through the analysis of variance with Trial (first versus second test trial) as a between subject factor and Sound (environmental sound versus speech sound) as a within subject factor Separate analyses were run for the iambic and trochaic cases. This way, the average of the distances was submitted to a 2(test) X 2(distance type: D_{es-i} or D_{sp-i}) ANOVA for the iambic case and (2(test) X 2(distance type: D_{es-t} or D sp-t) ANOVA for the trochaic case.

The analysis for the iambic tests revealed that after 100 epochs of exposure, the model showed no significant differences between environment and speech sound conditions ($F(1, 158) = 1.20, p = 0.39$). However, after more exposure to the input data the network showed greater familiarity (reduced distance) between speech and iambic test stimuli representations: ($F(1, 158) = 2.37, p = 0.0001$).

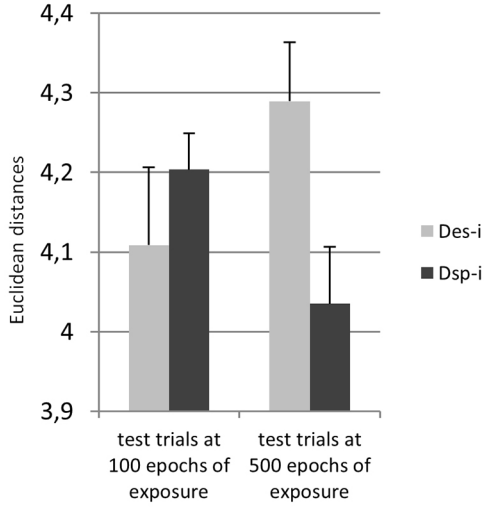

Figure 6.4: Average distances for 80 runs of the model. Light grey bars correspond to the mean distances between iambic test trials and environmental sounds representations (D_{es-i}). Dark grey bars correspond to the mean distances between iambic test trials and speech representations (D_{sp-i}). Error bars indicate 0.95 confidence interval of the standard error of the mean.

For trochaic test stimuli, the ANOVA revealed that in neither after 100 epochs of exposure ($F(1, 158) = 0.49, p = 0.49$), nor after 500 epochs of exposure ($F(1, 158) = 0.13, p = 0.71$) did the model show any significant differences in familiarity between the two types of test stimuli.

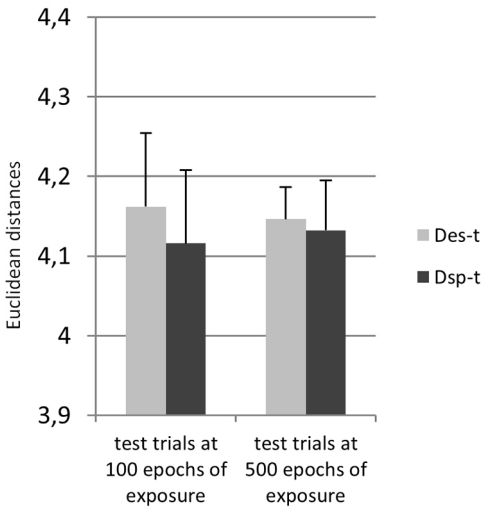

Figure 6.5: Average distances for 80 runs of the model. Light grey bars correspond to the mean distances between trochaic test trials and environmental sounds representations (D_{es-t}). Dark grey bars correspond to the mean distances between trochaic test trials and speech representations (D_{sp-t}). Error bars indicate 0.95 confidence interval of the standard error of the mean.

DISCUSSION

The model's developmental trajectory is analogous to that in the infant empirical data (Yoshida et al., 2010). In an initial exposure phase, both model representations for speech or environmental sounds did not differ in their averaged distances from the iambic test stimulus representation. In contrast, with additional exposure, the model develops opposite behaviours for each representation for speech and environmental sounds. For the environmental sounds' representation of the model, the distance relative to the iambic test trial representation, increased considerably. The distance between the model's representation of infant-directed speech relative to the iambic test trial representation, decreased considerably. This is not observable for trochaic stimuli (Figure 6.5). In fact, regardless of the results are statistically not relevant, the opposite behaviour can be observed, that is, D_{es-t} decreases and D_{sp-t} increases with exposure.

This means that the model treats the iambic test stimulus as more familiar to what it knows as infant-directed speech and, in addition, it organizes this test stimuli representation as different from the environmental sounds representation.

These results confirm our predictions, in which the model's exposure

to the durational patterns present in maternal speech and singing would influence, in a later stage, the construction and the organization of the representations for each stimulus, specifically the approximation of the speech and iambic durational sequence representations. Moreover, a transition from one early state into a later one is also observable in the model. That is, the model transforms, through the learning of speech, singing and environmental sounds, an early state of no familiarity with any type of durational sequences, either iambic or trochaic, into a later competence where an organization of the representations that approximates the iambic durational sequences to the speech data is observable.

However, it matters to discuss how this transition between two stages occurs. Furthermore, we have to examine the mechanisms by which the model acquired the later stage of development.

The developmental trajectory that was obtained might have been caused by several aspects. One possibility is the structure of the exposure data that shapes the model's representations. The fact that the nature of the data has influenced the construction of the internal representations reflects that the model extracted the durational patterns present in the data. Other aspect concerns how the model captures these patterns that are present in the data. There are two ways in which this could happen: (1) the nature of the algorithm, and (2) the encoding of the data.

When capturing patterns in a durational sequence (and specifically the iambic durational sequence where events are perceived as sequences of short-long events), it is essential to consider the relation between consecutive elements. We suggest that for processing duration patterns, not only the order of the events must be considered, but also the specific timing of each event. Moreover, the process of learning should simulate an exposure environment where there isn't a formal teacher. The positive modelling results suggest that the learning algorithm, encoding and training environment were all found to be suited, being validated by the achieved results.

Regarding the encoding options taken for handling the input data, these have demonstrated to be suited for the problem in question to solve, conveying to the model the notion of concurrence of duration, pitch, and loudness for each durational event and additionally the time dimension in the sequentially presentation of the durational events.

The nature of the environment that was recreated for the exposure in order to convey experience to the model, and how it influences the developmental trajectory followed is a subject of most interest. In this sense, we would like to investigate to which extent did the singing data biased the results of this simulation. Accordingly, we proceeded to the experiment 2.

6.4.3 Experiment 2: manipulating the exposure environment

RATIONALE

Recent research has shown correspondences in music and language processing and representation correspondences in adult humans. Patel et. al. (2006) have shown that the prosody of a culture's native language is reflected in the rhythms of its instrumental music. Language is also reflected in an individual's non-linguistic grouping preferences (Yoshida et al., 2010). Music has shown to be beneficial for speech development (Wibke et al., 2010). Together, these findings suggest that the processing and internal representations of music and language may interact fundamentally. Consequently, in this second experiment we aim to investigate the influence that the singing data has on the model's construction of the stimuli's representations. In other words, we ask if the model is not exposed to singing data then, does it produce a similar developmental trajectory as the one observed in the previous experiment?

In order to explore possible interactions between music and language, we propose an experiment in which the singing data is excluded. This way, we can observe if the same behaviour is produced as in the previous experiment or not. The total number of instances used as input data remain unchanged. The exposure environment will then be composed by 75% of speech instances (145) and 25% of environmental sounds instances (48). The experimental setup was the same as followed in the first experiment, performing two tests at different learning stages of the model in order to observe changes in representations over the course of exposure.

RESULTS

Figure 6.6, likewise Figure 6.4 illustrates the mean distances between internal representations of environmental sound and iambic and the mean distances between internal representations of speech and iambic representations, for two different points in learning. Similarly, Figure 6.7 illustrates the mean distances between environmental sound and trochaic sound representations and the mean distances between speech and trochaic representations for two different points in learning.

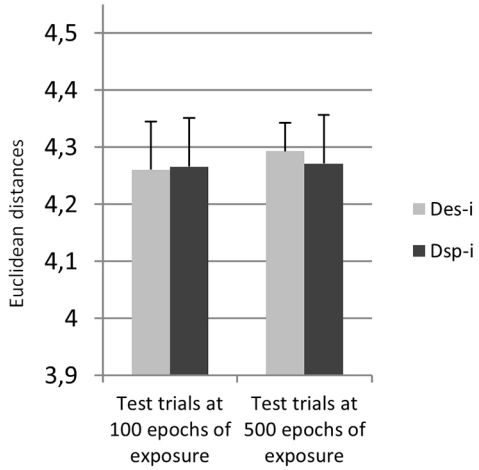

Figure 6.6: Average distances for 80 runs of the model that has not been exposed to singing data. Light grey bars correspond to the mean distances between iambic test trials and environmental sounds representations (D_{es-i}). Dark grey bars correspond to the mean distances between iambic test trials and speech representations (D_{sp-i}). Error bars indicate 0.95 confidence interval of the standard error of the mean.

The results were subjected to an analysis of variance with Trial (first test trial versus second test trial) as a between subject factor and Sound (environmental sound distance versus speech distance) as a within subject factor. Two separate ANOVAs were run for the iambic case and the trochaic case (2(test) X 2(distance type: D_{es-i} or D_{sp-i}) ANOVA for the iambic case and (2(test) X 2(distance type: D_{es-t} or D sp-t) ANOVA for the trochaic case).

The analysis reveals that either in the first test, after 100 epochs of exposure ($F(1, 158) = 0.008, p = 0.93$), or in the second test after 500 epochs of exposure($F(1, 158) = 0.19, p = 0.66$), the model had no significant different familiarity in the two distances measured.

In turn, ANOVA for trochaic test reveals that in either for the first test, after 100 epochs of exposure, or after more exposure, the model had no significant different familiarity in the two distances measured, respectively ($F(1, 158) = 0.0016, p = 0.97$), ($F(1, 158) = 1.6, p = 0.20$). In summary, no significant effects or interactions were found (all Fs < 1.5) in any of the analyses.

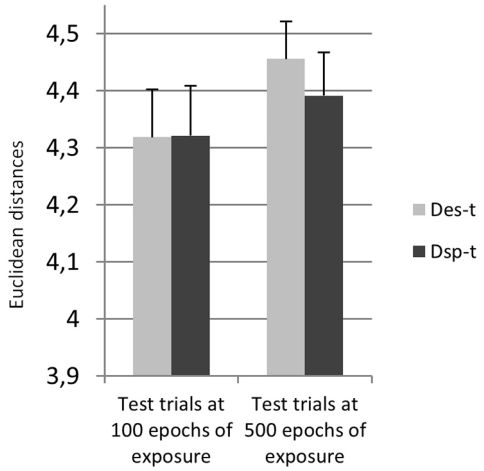

Figure 6.7: Average distances for 80 runs of the model that has not been exposed to singing data. Light grey bars correspond to the mean distances between trochaic test trials and environmental sounds representations (D_{es-t}). Dark grey bars correspond to the mean distances between trochaic test trials and speech representations (D_{sp-t}). Error bars indicate 0.95 confidence interval of the standard error of the mean.

DISCUSSION

These results confirm that the structure of the environment to which the model is exposed impacts dramatically on how the model encodes stimuli. In this specific case, the lack of infant directed singing in the exposure data influenced the way in which the model organized the representations. In this experiment, this becomes clear when obtaining different results from the experiment 1. When comparing Figure 6.6 that was obtained in experiment 2, with the Figure 6.4 from experiment 1, it is possible to observe the same pattern of behaviour, although in Figure 6.6 the results are statistically not significant. In the case of the environmental sounds – iambic, the distance is bigger in the older networks. In the case of the speech – iambic, the distances reveal an opposite track, where more experienced networks show smaller distances than the younger ones. The factor that is to point out between the two figures is that in the Figure 6.4 the behaviour noted before is magnified.

In particular, when comparing the results across the two experimental setups, they reveal that the infant-directed singing was essential in producing the outcome in the Experiment 1. Moreover, regarding the trochaic test, the results are opposite to the ones obtained in Figure 6.5, even though they are

not statistically relevant.

6.5 General discussion

The suggestion that the exposure to the surrounding sound environment influences the development of musical preferences and representations (Demorest et al., 2010) was explored by performing computer simulations with a computational model based on temporal information processing and representation. We observed the evolution of the internal representations of the computer model over time, throughout exposure to an auditory environment, and compared these representations with the representations resulting from the test stimuli.

The computer model is composed of two main modules. In the pre-processing module, the audio data is segmented and three dimensions of sound are extracted, namely pitch, duration and loudness. These elements are organized in such a way (see Figure 6.2) as to enable the handling pitch, duration and loudness in parallel for each durational event, while at the same time, maintaining the time information available in the sequential ordering of the durational events.

In the network module, the input data that comes from the pre-processing module is exposed to a neural network. The TD reinforcement leaning algorithm (Sutton, 1988) was used to train the network. This module is intended to develop knowledge of the exposure environment. Through exposure to the data, the network develops meaning or coherent internal representations. In the particular case, the knowledge that the model needs to acquire requires the processing of sound sequences, characterized by a succession of particular sounds that occur in specific orders (Warren, 1993). At some level of analysis, the comprehension of sequential structures such as speech or music (i.e., singing) involves the balance of different resolutions of examination, one at the level of the sequence of the components and other at the level of the global organization. This relates to mechanisms based on the global integration of items in sequences, such as grouping. The model must be able to capture the relation amongst successive events, as for a listener derives the rhythmic structure of a sequence of sounds during the course of speech or music processing. This involves two mechanisms that must operate concurrently: (i) order identification and preservation and; (ii) global recognition of overall patterns, such as an holistic pattern recognition, that does not imply the decomposition into component elements. It is possible that novice listeners are still not attending to this holistic dimension and are rather building it through experience (Warren, 1974).

In neural networks, knowledge representation can be interpreted as divided into a long-term memory of the network that is represented in the connection weights and a working memory that is represented in the transient activation patterns (Shultz, 2003). This way, previously acquired know-

ledge interacts with the current data in the interpretation of online auditory stimulation. But, like in human memory, new learning interferes slightly with old knowledge, but not catastrophically (Barnes & Underwood, 1959). Therefore, this characteristic of neural networks of slowly building long term memory, bit by bit with new information corroborates with the idea of building by exposure the holistic dimension for sequence processing.

The TD learning algorithm is centred on the idea of predicting a quantity that depends on future values of a given signal. Combined with reinforcement learning, the quantity to predict becomes the measure of the total amount of reward that is expected over the future. This way, the algorithm suits task that involve learning about temporal contingencies that span many time steps, and not just the immediately preceding context information. These characteristics have proven to be essential, along with the encoding of the input data, for capturing statistical properties of the data. The fact that the statistical properties that follow the iambic (short-long) patterns are reflected mostly at the phrasal level, according to the European Portuguese language's structural properties (Dryer & Haspelmath, 2011), the more global approach to the data that Temporal Differences enables, was fundamental for the model to become permeable to the properties of the data and contributed for the perspective on the problem that we seek that combines a detailed and a global dimension over the data.

When exposed to infant-directed speech, infant-directed singing and environmental sounds, the model revealed a developmental profile that is also observed in the experimental data (Yoshida et al., 2010). That is to say, initially the model did not show any specific organization of the representations for the test stimuli, neither the iambic nor the trochaic. In a later exposure phase, the model treated the representation for iambic test stimuli as similar to infant-directed speech and different from the other environmental sounds. In addition, when infant-directed singing was excluded, the model did not respond in the same manner, and failed to differentiate between test stimuli. But, what were the processes that operated in the model that led to these simulations results?

The way in which the model organized its internal representations was consistent with the statistical patterns of the language to which it was exposed. These results reveal the relevance of the exposure data on the process of building and organizing the internal representations of the model. We identified two factors that can contribute to the building and organization of the internal representations of the model: (i) the intrinsic characteristics of the structure of each element in the data and; (ii) the type of sound (i.e. infant-directed speech, infant-directed singing and environmental sounds) that made up the data set. Therefore, the importance is reflected on the statistical properties in each element of the exposure material and in the qualitative composition of the environment, specifically, whether if there is singing or not. The model is extremely sensitive to the statistical regularities that are present in the exposure environment and such regularities seem to

be extremely reflected on the singing training patterns.

In infant-directed singing, caregivers tend to use a particular style of singing that differs acoustically in a number of aspects from the typical adult singing style. In this special register, singing has slower tempo, relatively more energy at lower frequencies, lengthened inter-phrase pauses and has higher pitch and jitter factor (measure associated with increased intensity of emotion) (Trainor et al., 1997). These characteristics lead to a more accentuated articulation of words, when they are present. In addition, pitch variability is higher and the rhythm exaggerated in infant-directed play-songs, but not in lullabies (Trehub & Trainor, 1998). These acoustic modifications in infant directed singing not only attract infants' attention and may be used by adults to regulate infants states to communicate emotional information (Rock et al., 1999), but also seem to facilitate statistical learning.

It has previously been suggested that music can improve behavioural performance in various domains, including language (Moreno, 2011). Our results suggest that singing may also work as a facilitator in learning the temporal regularities of a language and of a culture. Moreover, if we assume that singing is a stimulus that is close to speech (in acoustic-physical terms i.e., speech with music on top) then music becomes the catalyst for learning. In other words, the addition of music makes the words better articulated, more rhythmically accented and, consequently, easier to parse, and learn about the language's rhythmic structure.

As a consequence of the simulations with the computer model, we have derived this hypothesis upon simulation data. Therefore, it would be interesting to test empirically with behavioural experiments if European Portuguese infants develop iambic grouping preferences and if the presence of singing in the auditory environment of babies biases in any way the acquisition of speech in the early development.

It is difficult to disclose how development and learning interact. We assume that development occurs through some kind of learning of long-term memory elements (Shultz, 2003). But this assumption does not totally clear up whether a task requires learning time or restructuring knowledge representations (i.e., development) of certain competences.

Development can be considered as a process through which an early state is transformed into a later one (see Sections 2.3.1 and 3.1.3). Driven by experience, it is observable that a new competence is acquired in the later stage. During the simulations performed, we have observed the progression of the internal representations of the model. We have observed that the model's connections strengths changed through experience, showing an adapting cognitive structure with plasticity behaviour. We have analysed its activation patterns changing through a learning process. These representations evolved from an initial stage where there is no observable familiarity with any type of durational sequences, either iambic or trochaic, to a later one where an organization of representations that approximates the iambic

durational sequences to the speech data. Therefore, in the sense that the model's weighted connections were shaped by continuous adaptations to the data to which it is exposed, and as a result, qualitative different representations emerged, reflected in the progress that the model achieves in encoding the input data, the model can be considered developmental.

However, such changes brought about by weight adjustments and activation patterns are quantitative, made within the network's current typology. Considering that cognitive development implies qualitative changes in the structure that supports cognition, that is, changes in the existing structure to allow more complex processing and provide the enabling conditions that allow learning to be most effective (Elman, 1993), we cannot argue that the network's processing structure is not qualitatively different than before. Following this framework for development, it is more likely that the model has no developmental attributes and the transitions between stages are just the effect of a slow learning environmental-shaped process. However, despite the static network architecture, it is undeniable that the model was able to build, driven by experience, a representation for the exposure stimuli, as a long-term memory. This representation, which can be seen as a cognitive structure for interpreting the environment, influenced the different ways that the model encoded the test stimuli, in an initial and a later stage.

The building of these representations becomes, then, cultural dependent. This environmental constraint that is shaped by culture might be conditioning our later appreciation of music (Hannon & Trehub, 2005). The adaptive advantages of learning, driven by culture, provide a non-genetically based transmission of behaviours, shape predispositions in a similar way within the people of one culture and, therefore, promoting social cohesion (Cross, 2001). Our modelling results are consistent with the idea that music, like speech, is an integral dimension of human development.

To conclude, with the computer model, we attempt to capture the nature of the transition mechanisms that can account for how one level of performance is transformed into the next level of performance. This way, computer model simulations provided a useful tool for discussing and reflecting about the causal mechanisms involved in the development of rhythmic representations and predispositions as a result of exposure to the temporal prosodic patterns of a specific culture. The computational approach additionally enabled going beyond the laboratory stimuli, complementing experimental data and suggesting new directions for future empirical investigations.

6.6 Conclusions

In the research work reported, we built a computational model based on temporal information processing and representation aiming to investigate how the temporal prosodic patterns of a specific culture influence the development of rhythmic representations and predispositions. Additionally, we

explore possible interactions between music and language at an early age. With this purpose, the research was organized into two simulations.

In the first simulation, we have observed a developmental behaviour captured by the transition from one early state into a later one, which reflects the influence of the amount of experience of the model on the construction and organization of its internal representations. These representations show that the iambic test stimulus is familiar to what the model knows as infant-directed speech and, in addition, organizes this test stimuli representation as a different from the environmental sounds representation, as it was hypothesised. Consequently, our first simulation verifies the predictions made and shows that the model provides a reasonable account of the behavioural data.

In an effort to understand the mechanisms that underlie the developmental behaviour found, the simulation generated a new problem, in particular, which was the influence of the singing data on the model's construction of the stimuli's representations. For this reason, the model was submitted to another experiment, aiming to observe how it would respond in a novel situation. The results demonstrated that the lack of infant-directed singing in the exposure data influenced the way in which the model organized the representations, verifying that the structure of the exposure environment to which the model is exposed has impact on how the model encodes stimuli.

As a final note, these conclusions do not intend to claim that the model is an accurate reproduction of the specific mechanisms operating in human infants. Rather, we claim that there is a correspondence between the behaviours and, therefore, between the kinds of constraints that are built into the model that might operate in humans (Marr, 1982). For this reason, we consider that the model captures the essence of what is going on in the human mechanisms dealing with speech and music rhythm computation.

7

Conclusions

We have explored computational solutions suitable to each specific research stage, that can best contribute to the study of the way human cognitive structure is shaped to build musical predispositions during early development. We have also explored a comparative approach to the study of early infant development of musical predispositions by searching for possible interactions and correlations involving music and language.

We have started from the hypothesis that the development of musical representations and predispositions in infants is influenced by the prosodic features that are present in the sonic environment of their culture. This hypothesis, led to a further question: how would such features or elements influence the development of musical predispositions and representations during infancy?

We conclude, from the prosodic characterization performed, that it reinforces the notion that the melodic information present in the prosody of speech contains information that can be used to distinguish communicative contexts from one another (Fernald, 1989; Papoušek et al., 1990; Papoušek & Papoušek, 1991). Vowels contain the melodic information present in speech and, in this sense, they carry the intentional value, the emotional content of communication.

The experiments carried out suggest that the processing of speech and singing may share the analysis of the same properties of the stimuli, given that common features were used by the classification model to discriminate speech and singing. These results strengthen previous findings by providing further evidences that the cognition of both music and language during the pre-verbal period might share computational resources (Patel, 2008).

Furthermore, we have observed relevant rhythm differences between infant-directed speech and songs from the two Portuguese variants, but not melodic differences. This shows that despite caregivers' cross-cultural adjustments in their vocal interactions with infants, a rhythmic cultural identity is kept in speech and vocal songs. This reinforces the hypothesis of a relation between the rhythms of music and language in the context a particular culture (Patel, 2006; Hannon, 2009).

The building of musical representations and predispositions, specifically temporal representations, is dependent on the auditory environment. It is from the exposure to the structure of the sonic background that we bias the formation of the cognitive structure that supports our understanding of musical experience. Among the auditory stimuli that surround us, the structure of language has a predominant role in the building of musical predispositions and representations.

The process through which pre-verbal infants become perceptually tuned into their native language influences as well their musical sound perception system and contributes to shape their cognitive structure. Hence, infants gain sensitivity to culture-specific organizational principles and become acculturated listeners.

This auditory environmental bias might be related to two main factors. These are the statistical properties present in the auditory environment and the qualitative composition of the environment, specifically whether there is singing or not. Moreover, the lack of vocal songs might compromise the transmission of a rhythmic cultural identity, since singing might work as a facilitator in learning the temporal regularities of a language and of a culture.

Concerning the mechanisms that drive the biasing acquired through the auditory environment, we consider that the comprehension of sequential structures such as speech or music involves the balance of different levels of cognitive processing, one at the level of the sequence of components and other at the level of global organization. Temporal Differences learning algorithm (one of the computational modelling techniques used in our experiments) combines a detailed and a global dimension over the data.

There is a reciprocal influence between music and language. They interact both ways as music, too, has an impact on language. Music is responsible for imprinting emotional meaning into speech. It is through the vowels, the musical channel present in speech, that emotion is transmitted. This creates a paralinguistic meaning beyond the words content, which is carried in the musical layer of speech. This is in accordance with the theory that defends that music is a natural sign of emotions with social functions in communication (Thomas, 1995). This emotional meaning that is dragged by music into speech becomes crucial for communication with infants, who still have no knowledge of verbal content.

The rhythm present in language and vocal songs is a key element in characterizing the identity of a culture. It is through the exposure to the structure of native language and vocal songs that rhythmic preferences and representations are shaped. Vocal songs play a paramount role in the process of transmitting the rhythmic identity of a culture. Their lack, indeed, might even compromise the learning of the rhythmic structure of the native language, and thus the building of the rhythmic representations and preferences characteristic of a given culture. The building of these representations becomes, thus, cultural determined. This culturally shaped environmental

constraint might later condition our appreciation of music (Hannon & Trehub, 2005).

The adaptive, culture-driven advantage of learning, shapes predispositions in a similar way within people of a given culture and consequently promotes social cohesion (Cross, 2001). This dissertation bears out the idea that music, like speech, is a product of both social interaction along with biology and a necessary and integral trait of human development.

Assuming that music can influence both language learning and the transmission of a cultural rhythmic identity, we consider music a "transformative technology" (Patel, 2008). Although music might be built on existing processing mechanisms serving other capacities such as language it transforms our brain systems and thus our experience of the world.

7.1 Summary of contributions

The outcomes of this dissertation include practical contributions such as a database, a computational model and an extensive acoustic characterization of infant directed speech and songs from two Portuguese variants. Additionally, we consider relevant the theoretical contributions in which we include the knowledge gained, the additional evidence for existent hypothesis and the additional support for the methods followed and predictions that were generated.

We present them next.

A Database containing infant-directed speech and singing. The database is composed of samples from two Portuguese variants, namely European and Brazilian Portuguese, and different care giving contexts, comprising a total number of 977 instances. (See section 5.2.1 Corpus.)

A computational model. The model was developed and evaluated, providing a useful tool for exploring and reflecting about the causal mechanisms involved in the development of rhythmic representations and predispositions. The model can be further used as a framework for exploring possible interactions and parallels between music and language.

An extensive acoustic characterization of infant-directed speech and singing. This characterization considered European and Brazilian Portuguese and captured rhythmic and melodic patterning. The characterization reveals that there are relevant rhythm differences between the two Portuguese variants but not melodic differences, which shows that rhythm features are a key element for characterizing the prosody from language variants and songs from each culture.

Additional findings for existing hypothesis. These findings support present theories of music and language' sharing of processing resources (Patel, 2008) and its mutual interactions. Specifically, in Chapter 5, where results indicate that the processing of speech and singing may share the analysis of the same stimulus' properties, given that common features were used by the classification model to discriminate speech and singing. Moreover, in chapter 6, vocal songs' material is shown to be a facilitator in the model's learning of the temporal regularities of the language.

Support for the validity of techniques followed. This support is traduced into further research work concerning computational modelling as a means of studying cognitive phenomena specifically research into music cognition from a comparative perspective that includes its interactions with language.

New hypotheses were generated on the model's predictions. This process was a consequence of the computational modelling methodology used. The hypotheses provide a solid basis for further refinement and exploration. Furthermore, they provide consistent new possibilities for future empirical research, to be tested through behavioural experimentation. Specifically, suggestions are given towards testing Portuguese infants for their rhythmic grouping preferences in different stages of development. Additionally, we propose to test the influence of the presence of vocal songs in infants' auditory environment in the acquisition of language, during early development.

Knowledge gained. Further insights were obtained by searching for factors specifically, language and its rhythmic structure that contribute to shape our cognitive structure and influence the cognition of music by building and shaping our predispositions and representations.

7.2 Future directions

Despite what we have accomplished so far, we consider this research work an early approach towards the identification of the factors that contribute to shape of our musical predispositions. Much has been left to explore. Accordingly, we identify, in one hand, experiments that have not been performed mostly for time constraints and could have enriched this research; and also, lines of research that we see as promising for the future. These include:

- After a computational model is implemented, tested and validated, hundreds of possibilities arise for experiments that could be performed. In the case of the model based on temporal processing, we find that testing other languages, in addition to the experiments that we have

carried out, could turn out to be interesting and enriching. Brazilian Portuguese - a language with approximately the same lexicon as European Portuguese but with structural rhythmic differences - would provide the opportunity to test the exposure to them and the resulting influence of this exposure on the building of representations. Testing the model with Japanese would provide the opportunity to observe the model's behaviour with a language that is completely different from Portuguese in its structure. Additional validation would be thus provided.

Simulations could also be performed with a view to identifying a hypothetical critical period wherein exposure to singing has an effect, in the sense of facilitation, on the learning of the rhythmic structure of the native language. This could be done by varying the time step in which singing materials are introduced in the training data.

- Regarding the model, we find generative algorithms (Fahlman, 1990) very appealing for modelling the developmental phenomena of music cognition. Constructivist networks integrate neural cognitive and computational perspectives that result in models whose architecture grows by hidden-unit recruitment, changing the existent processing structure (Mareschal & Shultz, 1996). This way, constructivist models can draw a clearer distinction between learning and development, where learning is represented by the quantitative change in the parameter values within the network and development is the qualitative change in the structure of the model (Shultz et al., 1995).

- We consider that for the cognition of music, audition is not the only physical information input. Other elements contribute to it, such as vision (Cvejic et al., 2012), and even the whole body, which, when exposed to sound, functions through haptic perception, as a resonance cavity and stores the memory of its experiences and of proprioceptory information (Leman, 2008). Computer algorithms that aim modelling music cognition must consider these elements and include a connection to the world through a body. One way of doing so is to embed algorithms in a body by means of robotics. In developmental robotics, cognitive computational models are embodied, allowing the models to interact and experience the real world (Lungarella et al., 2003). This concept of embodied cognition modelling, where a body mediates perception, would allow to articulate music cognition models with the perceptual variables referred above. This approach would bring about advances in the identification of the factors affecting our music cognition and provide, in addition, a better understanding of the way they interact.

- Music technologies are built for people. They provide tools for sharing, discovering and making music, among many other purposes. Meanwhile, researchers are seeking for ways of improving the implementa-

tion of the mechanisms involved in the building of these tools, such as semantic analysis of sound or music and sound separation. However, the implementations of these mechanisms have still barely matched human performance (Guaus & Herrera, 2006). Concomitantly, there is still little insight into how human musical cognitive performance works and how the human brain functions. Computational modelling can be a valuable contribution towards driving forward the understanding of the mechanisms that underlie music cognition. It is our conviction that the more knowledge is produced about human musical cognition, and the more these findings are applied to the building of music technologies, the better the tools will perform. Building music technologies grounded in solid musical cognitive findings will push algorithms to a closer approximation of human performance and, therefore, produce tools that better serve peoples' needs. On the other hand, this knowledge, once produced, can also become a synergy for building more apt computational models that will contribute, in their turn, to the further study of music cognitive mechanisms.

Bibliography

Ahmad, K., Casey, M., & Bale, T. (2002). Connectionist simulation of quantification skills. *Connection Science*, *14*(3), 165–201.

Anderson, j. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Baker, M. (1989). A computational approach to modeling musical grouping structure. *Contemporary Music Review*, *4*, 311–325.

Barbosa, P. A. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication*, *49*, 725–742.

Barbosa, P. A. & Bailly, G. (1994). Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, *15*((1-2)), 127–137.

Barnes, J. & Underwood, B. (1959). "fate" of first-list associations in transfer theory. *Journal of experimental psychology*, *58*(2), 1–97.

Bates, E., Thal, D., Finlay, B., & Clancy, B. (2002). *Handbook of Neuropsychology, 2nd Edition, Vol. 8, Part II*, chap. Early language development and its neural correlates, pp. 525–592. Elsevier.

Bechtel, W. & Graham, G. (Eds.) (1998). *A companion to cognitive science.* Cambridge, UK: Blackwell.

Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: Rhythmical basis of speech representations in neonates. *Language and Speech*, *38*(4), 311–329.

Besson, M. & Schön, D. (2001). Comparison between language and music. *Annals of the New York Academy of Sciences*, *930*(1), 232–258.

Biggs, J. & Collis, K. (1982). *Evaluating the quality of learning*. New York: Academic Press New York.

Bispham, J. (2006). Rhythm in music: What is it? who has it? and why? *Music Perception: An Interdisciplinary Journal*, *24*(2), 125–134. Papers from the 10th Rhythm Perception and Production Workshop.

Bolton, T. (1894). Rhythm. *American Journal of Psychology*, *6*, 145–238.

Bomba, P. & Siqueland, E. (1983). The nature and structure of infant form categories. *Journal of Experimental Child Psychology*, *35*(2), 294–328.

Brandt, A., Gebrian, M., & Slevc, L. (2012). Music and early language acquisition. *Frontiers in Psychology*, *3*, 1–17.

Bregman, A. (1994). *Auditory scene analysis: The perceptual organization of sound.* The MIT Press.

Bryant, G. & Barrett, H. (2007). Recognizing intentions in infant-directed speech evidence for universals. *Psychological Science*, *18*(8), 746–751.

Callan, D., Tsytsarev, V., Hanakawa, T., Callan, A., Katsuhara, M., Fukuyama, H., Turner, R. et al. (2006). Song and speech: brain regions involved with perception and covert production. *Neuroimage*, *31*(3), 1327–1342.

Cangelosi, A. (2005). The emergence of language: neural and adaptive agent models. *Connection Science*, *17*(3), 185–190.

Carroll, S. (2003). Genetics and the making of homo sapiens. *Nature*, *422*(6934), 849–857.

Carterette, E. C. & Kendall, R. A. (1999). *The psychology of music, second edition*, chap. Comparative music perception and cognition, pp. 725–791. Academic Press San Diego, CA.

Cheng, Y., Lee, S., Chen, H., Wang, P., & Decety, J. (2012). Voice and emotion processing in the human neonatal brain. *Journal of Cognitive Neuroscience*, *24*(6), 1411–1419.

Christiansen, M. H. & Chater, N. (Eds.) (2001). *Connectionist Psycholinguistics.* Westport, CT: Ablex.

Clarkson, M. G. & Clifton, R. K. (1984). Infant pitch perception: Evidence for responding to pitch categories and the missing fundamental. *Journal of the Acoustical Society of America*, *77*(4), 1521–1528.

Clarkson, M. G., Martin, R. L., & Miciek, S. (1996). Infants perception of pitch: Number of harmonics. *Infant Behaviour and Development*, *19*, 191–197.

Cook, N. (1992). *Music, imagination, and culture.* Oxford University Press, USA.

Cooper, R. & Aslin, R. (1990). Preference for infant-directed speech in the first month after birth. *Child development*, *61*(5), 1584–1595.

Cooper, R. & Aslin, R. (1994). Developmental differences in infant attention to the spectral properties of infant-directed speech. *Child Development*, *65*(6), 1663–1677.

Cross, I. (2001). Music, cognition, culture and evolution. *Annals of the New York Academy of Sciences*, *930*, 28–42.

Cross, I. (2012). *Language and Music as Cognitive Systems*, chap. Music as a social and cognitive process, pp. 315–328. Oxford: Oxford University Press.

Cross, I. & Tolbert, E. (2008). *The oxford handbook of music psychology*, chap. Music and meaning, pp. 24–34. Oxford University Press.

Cvejic, E., Kim, J., & Davis, C. (2012). Recognizing prosody across modalities, face areas and speakers: Examining perceivers sensitivity to variable realizations of visual prosody. *Cognition*, *122*(3), 442–453.

Dalla Bella, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition*, *80*, B1–B10.

Darwin, C. (1871). *The descent of man, and selection in relation to sex.* London: John Murray.

Demorest, S., Morrison, S., Stambaugh, L., Beken, M., Richards, T., & Johnson, C. (2010). An fmri investigation of the cultural specificity of music memory. *Social cognitive and affective neuroscience*, *5*(2-3), 282–291.

Desain, P. & Honing, H. (1991). *Music and connectionism*, chap. The quantization of musical time: A connectionist approach., pp. 150–160. Cambridge: The MIT Press.

Desain, P. & Honing, H. (1999). Computational models of beat induction: The rule-based approach. *Journal of New Music Research*, *28*(1), 29–42.

Deutsch, D. (1991). The tritone paradox: An influence of language on music perception. *Music Perception*, *8*, 335–347.

Deutsch, D., Henthorn, T., & Dolson, M. (2004). Absolute pitch, speech, and tone language: Some experiments and a proposed framework. *Music Perception*, *21*(3), 339–356.

Dowling, W. (1999). *The psychology of music, second edition*, chap. The development of music perception and cognition, pp. 603–626. Academic Press San Diego, CA.

Dowling, W. & Harwood, D. (1986). *Music cognition*, vol. 19986. Orlando: Academic Press.

Dryer, M. S. & Haspelmath, M. (Eds.) (2011). *The World Atlas of Language Structures Online.* Munich: Max Planck Digital Library.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. L. (2005). Connectionist models of cognitive development: where next? *Trends in Cognitive Sciences*, *9*(3), 111 – 117. Developmental cognitive neuroscience.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development.* Cambridge, MA: The MIT Press.

Ettlinger, M., Margulis, E., & Wong, P. (2011). Implicit memory in music and language. *Frontiers in Psychology*, *2*, 1–10.

Fahlman, S. E. (1990). *Advances in neural information processing systems 2*, chap. The cascade-correlation learning architecture, pp. 524–532. Los Altos, CA: Morgan Kaufmann.

Falk, D. (2004). Prelinguistic evolution in early homnids: whence motherese? *Behavioral and Brain Sciences*, *27*, 491–503.

Feldman, D. H. (1980). *Beyond universals in congnitive development.* Norwood, NJ: Ablex.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant behavior and development*, *8*(2), 181–195.

Fernald, A. (1989). Intonation and communicative intent in mother's speech to infants: is the melody the message? *Child Development*, *60*, 1497–1510.

Fernald, A. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*, chap. Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective., pp. 391–428. London: Oxford University Press.

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages, , 64 (3), . *Child Development*, *64*(3), 657–674.

Fischer, K., Pipp, S. L., & Bullock, D. (1984). *Continuities and discontinuities in development*, chap. Detecting discontinuities in development: Methods and measurements, pp. 95–121. Norwood, NJ: Ablex.

Fischer, K. & Silvern, L. (1985). Stages and individual differences in cognitive development. *Annual Review of Psychology*, *36*(1), 613–648.

Fitch, W. (2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100*(1), 173–215.

Fodor, J. A. (1983). *The Modularity of Mind*. The MIT Press.

Fraisse, P. (1984). Perception and estimation of time. *Annual Review of Psychology*, *35*(1), 1–37.

Frota, S. & Vigario, M. (2001). On the correlates of rhythmic distinctions: the european/brazilian portuguese case. *Probus*, *13*(2), 247–275.

Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In *Proceedings of Prosody 2002*, pp. 323–326. Aix-en-Provence.

Gasser, M., Eck, D., & Port, R. (1999). Meter as mechanism: A neural network model that learns metrical patterns. *Connection Science*, *11*(2), 187–216.

Gentner, D., Rattermann, M., Markman, A., & Kotovsky, L. (1995). *Developing cognitive competence: New approaches to process modeling*, chap. Two forces in the development of relational similarity, pp. 263–313. Hillsdale, NJ: LEA.

Gilbert, N. & Troitzsch, K. G. (2005). *Simulation for the Social Scientist.*, chap. Simulation as a Method, pp. 15–27. Open University Press.

Goldman-Rakic, P. (1987). Development of cortical circuitry and cognitive function. *Child development*, *58*, 601–622.

Gollin, E. S. (1984). Early experience and developmental plasticity. *Annals of Child Development*, *1*, 239–261.

Gordon, W. (1989). *Learning and memory.* Pacific Greove, CA: Thomson Brooks/Cole Publishing Co.

Gouyon, F. (2005). *A computational approach to rhythm edscription.* Ph.D. thesis, Universitat Pompeu Fabra.

Grabe, E. & Low, E. L. (2002). *Laboratory Phonology 7*, chap. Durational variability in speech and the rhythm class hypothesis., pp. 515–546. Berlin, Germany: Mouton de Gruyter.

Graven, S. & Browne, J. (2008). Auditory development in the fetus and infant. *Newborn and Infant Nursing Reviews*, *8*(4), 187–193.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). *The cambridge handbook of computational psychology*, chap. Bayesian models of cognition, pp. 59–100. Cambridge University Press.

Guaus, E. & Herrera, P. (2006). Music genre categorization in humans and machines. In *121th AES Convention.*

Hall, D. (1991). *MusicalAcoustics, 2nd Edition.* Pacific-Grove,CA:Brooks/Cole Publishers.

Hannon, E. E. (2009). Perceiving speech rhythm in music: Listeners classify instrumental songs according to language of origin,. *Cognition*, *111*(3), 403–409.

Hannon, E. E. & Trehub, S. E. (2005). Metrical categories in infancy and adulthood. *Psychological Science*, *16*, 48–55.

Hargreaves, D. (1986). *The developmental psychology of music.* Cambridge University Press.

Hauser, M. & McDermott, J. (2003). The evolution of the music faculty: A comparative perspective. *Nature neuroscience*, *6*(7), 663–668.

Hawkins, J. & Blakeslee, S. (2005). *On intelligence.* Owl Books.

Haykin, S. (2009). *Neural networks and learning machines, Third Edition.* New Jersey: Pearson Prentice Hall.

Hazan, A., Holonowicz, P., Salselas, I., Herrera, P., Purwins, H., Knast, A., & Durrant, S. (2008). Modeling the acquisition of statistical regularities in tone sequences. In *30th Annual Meeting of the Cognitive Science Society.*

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological approach.* New York: John Wiley & Sons.

Hevner, K. (1935). The affective character of the major and minor modes in music. *The American Journal of Psychology*, *47*(1), 103–118.

Hinton, G. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, vol. 1, p. 12. Amherst, MA.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, *40*, 185–234.

Honing, H. (2005). Is there a perception based alternative to kinetic models of tempo rubato? *Music Perception*, *23*(1), 79–85.

Honing, H. (2006). Computational modeling of music cognition: A case study on model selection. *Music Perception*, *23*(5), 365–376.

Honing, H. & Ploeger, A. (2012). Cognition and the evolution of music: Pitfalls and prospects. *Topics in Cognitive Science*, *1*, 1–12.

Houghton, G. (2005). *Connectionist models in cognitive psychology.* Hove, Sussex, UK: Psychology Press.

Houston-Price, C. & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*, 341348.

Hukin, R. & Darwin, C. (1995). Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. *Attention, Perception, & Psychophysics*, *57*(2), 191–196.

Hunter, M. A. & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, *5*, 69–95.

Huron, D. (2001). Is music an evolutionary adaptation? *Annals of the New York Academy of Sciences*, *930*(1), 43–61.

Ilari, B. (2002). Music perception and cognition in the first year of life. *Early Child Development and Care*, *172*(3), 311–322.

Iversen, J. R., Patel, A. D., & Ohgushi, K. (2008). Perception of rhythmic grouping depends on auditory experience. *The Journal of the Acoustical Society of America*, *124*(4), 2263–2271.

Jackendoff, R. & Lerdahl, F. (2006). The capacity for music: What is it, and what's special about it? *Cognition*, *100*(1), 33 – 72. The Nature of Music.

Johnson, M. H. & Hann, M. (2011). *Developmental cognitive neuroscience, Third edition.* Wilwy-Blackwell.

Jusczyk, P., Luce, P., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.

Juslin, P. & Sloboda, J. (Eds.) (2001). *Music and emotion: theory and research.* Oxford ; New York : Oxford University Press.

Katz, G. S., Cohn, J. F., & Moore, C. A. (2008). A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Development*, *67*(1), 205 – 217.

Kemler, N. D., Jusczyk, P., Mandel, D., Myers, J., Turk, A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, *18*, 111–116.

Klahr, D. (1984). *Mechanisms of cognitive development.*, chap. Transition processes in quantitative development., pp. 101–139. New York: Freeman.

Knudsen, E. (2004). Sensitive periods in the development of the brain and behavior. *Journal of Cognitive Neuroscience*, *16*(8), 1412–1425.

Koelsch, S. (2012). *Brain and music*. Wiley-Blackwell.

Kotilahti, K., Nissilä, I., Näsi, T., Lipiäinen, L., Noponen, T., Meriläinen, P., Huotilainen, M., & Fellman, V. (2010). Hemodynamic responses to speech and music in newborn infants. *Human brain mapping*, *31*(4), 595–603.

Kraus, N. & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*(8), 599–605.

Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, *25*(1), 161–168.

Kruschke, J. K. (2008). *The cambridge handbook of computational psychology*, chap. Models of categorization, pp. 267–301. New York, USA: Cambridge University Press.

Kuhl, P. (2000). *The New Cognitive Neurosciences*, chap. Language, mind and brain: experience alters perception, pp. 99–115. The MIT Press.

Laske, O. (1988). Can we formalize and program musical knowledge? an inquiry into the focus and scope of cognitive musicology. *Musikometrika*, *1*, 257–280.

Leman, M. (2008). *Embodied music cognition and mediation technology*. The MIT Press.

Lerdahl, F. & Jackendoff, R. (1983). *A generative theory of tonal music*. MA: The MIT Press.

Lerner, R. M. (1984). *On the nature of human plasticity*. New York: Cambridge University Press.

Levitin, D. & Rogers, S. (2005). Absolute pitch: perception, coding, and controversies. *Trends in Cognitive Sciences*, *9*(1), 26–33.

Lindner, K. & Hohenbergen, A. (2009). Introduction: concepts of development, learning and acquisition. *Linguistics*, *47*(2), 221–239.

Ling, L. E., Grabe, E., & Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in singapore english. *Language and Speech*, *43*(4), 377–402.

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection Science*, *15*(4), 151–190.

Lynch, M., Eilers, R., Oller, D., & Urbano, R. (1990). Innateness, experience, and music perception. *Psychological Science*, *1*(4), 272–276.

MacWhinney, B. (2000). The childes project: Tools for analysing talk. third edition. mahwah, nj: Lawrence erlbaum associates.

Mahdhaoui, A., Chetouani, M., Zong, C., Cassel, R., Saint-Georges, C., Laznik, M.-C., Maestro, S., Apicella, F., Muratori, F., & Cohen, D. (2009). Automatic motherese detection for face-to-face interaction analysis. In A. Esposito, A. Hussain, M. Marinaro, & R. Martone (Eds.) *Multimodal Signals: Cognitive and Algorithmic Issues*, *Lecture Notes in Computer Science*, vol. 5398, pp. 248–255. Springer Berlin / Heidelberg.

Mandler, J. (1992). How to build a baby: Ii. conceptual primitives. *Psychological review*, *99*(4), 587–604.

Mandler, J. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognition and Development*, *1*(1), 3–36.

Marechal, D. & French, R. (2000). Mechanisms of categorization in infancy. *Infancy*, *1*(1), 59–76.

Mareschal, D., French, R., & Quinn, P. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, *36*(5), 635–645.

Mareschal, D., Johnson, M., Sirois, S., Spratling, M., Thomas, M., & Westermann, G. (2007). *Neuroconstructivism, Vol. I: How the brain constructs cognition.* Oxford, UK: Oxford University Press.

Mareschal, D. & Quinn, P. (2001). Categorization in infancy. *Trends in Cognitive Sciences*, *5*(10), 443–450.

Mareschal, D. & Shultz, T. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development*, *11*(4), 571–603.

Mareschal, D. & Thomas, M. S. (2006). How computational models help explain the origins of reasoning. *IEEE Computational Intelligence Magazine*, *1*(3), 32–40.

Mareschal, D. & Thomas, M. S. C. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation*, *11*(2), 1–14.

Marr, D. (1982). *Vision : A computational investigation into the human representation and processing of visual information.* WH Freeman.

Masataka, N. (1999). Preference for infant-directed singing in 2-day-old hearing infants of deaf parents. *Developmental Psychology*, *35*(4), 1001.

Masataka, N. (2006). Preference for consonance over dissonance by hearing newborns of deaf parents and of hearing parents. *Developmental science*, *9*(1), 46–50.

Masataka, N. (2009). The origins of language and the evolution of music: A comparative perspective. *Physics of Life Reviews*, *6*(1), 11 – 22.

McClelland, J. L. & Jenkins, E. (1991). *Architectures for Intelligence*, chap. Nature, nurture, and connections: Implications of connectionist models for cognitive development., pp. 41–73. LEA, Hillsdale, NJ.

McDermott, J. & Hauser, M. (2005). The origins of music: Innateness, uniqueness, and evolution. *Music Perception*, *23*(1), 29–59.

McDermott, J. & Oxenham, A. (2008). Music perception, pitch, and the auditory system. *Current opinion in neurobiology*, *18*(4), 452–463.

McMullen, E. & Saffran, J. (2004). Music and language: a developmental comparison. *Music Perception*, *21*(3), 289–311.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). *Signal to syntax: bootstrapping from speech to grammar in early acquisition.*, chap. Coping with linguistic diversity: The infant's viewpoint., pp. 101–116. Mahwah NJ: Lawrence Erlbaum Associates.

Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., & Friederici, A. (2012). Perception of words and pitch patterns in song and speech. *Frontiers in Psychology*, *3*(76), 1–13.

Mertens, P. (2004). The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In B. Bel & I. Marlien (Eds.) *Proc. of Speech Prosody*. japan.

Meyer, L. (1956). *Emotion and meaning in music*. Chicago, IL: University of Chicago Press.

Moore, J. & Guan, Y. (2001). Cytoarchitectural and axonal maturation in human auditory cortex. *Journal of the Association for Research in Otolaryngology*, *2*(4), 297–311.

Moreno, S. A. (2011). Short-term music training enhances verbal intelligence and executive function. *Psychological Science*, *22*(11), 1425–1433.

Morillon, B., Lehongre, K., Frackowiak, R., Ducorps, A., Kleinschmidt, A., Poeppel, D., & Giraud, A. (2010). Neurophysiological origin of human brain asymmetry for speech and language. *Proceedings of the National Academy of Sciences*, *107*(43), 18688–18693.

Morrison, S., Demorest, S., Aylward, E., & Cramer, S. (2003). Fmri investigation of cross-cultural music comprehension. . *Neuroimage , 20, 378-384.*., *20*, 378–384.

Morrison, S. J. & Demorest, S. M. (2009). Cultural constraints on music perception and cognition. *Progress in Brain Research*, *178*, 67–77.

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: a pdp model of the ab task. *Developmental Science*, *1*, 161–186.

Munakata, Y. (2004). Computational cognitive neuroscience of early memory development. *Developmental Review*, *24*(1), 133–153.

Munakata, Y. & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*(4), 413–429.

Nakata, T. & Trehub, S. (2004). Nakata, t., & trehub, s. (2004). infants responsiveness to maternal speech and singing. infant behavior and development , 27 (4),. *Infant Behavior and Development*, *27*(4), 455–464.

Nazzi, T. & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, *41*, 233–243.

Nespor, M., Shukla, M., van de Vijver, R., Avesani, C., Schraudolf, H., & Donati, C. (2008). Different phrasal prominence realizations in vo and ov languages. *Lingue e linguaggio*, *7*(2), 139–168.

Nettl, B. (2000). *The origins of music*, chap. An ethnomusicologist contemplates universals in musical sound and musical culture, pp. 463–472. MA: Massachusetts Institute of Technology Cambridge.

Nettl, B. (2005). *The study of ethnomusicology: thirty-one issues and concepts*. University of Illinois Press.

Newell, A. (1973). *Visual Information Processing*, chap. Production systems: Models of control structures., pp. 463–526. San Diego, CA: Academic Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A. & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126.

Nooteboom, S. (1997). *The handbook of phonetic sciences*, chap. The prosody of speech: melody and rhythm, p. 640673. Cambridge, MA: Blackwell.

Olsho, L. W., Schoon, C., Sakai, R., Turpin, R., & Sperduto, V. (1982). Preliminary data on frequency discrimination in infancy. *Acoustical Society of America*, *71*, 509–511.

O'Reilly, R. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in cognitive sciences*, *2*(11), 455–462.

O'Reilly, R. C. & Munataka, Y. (2000). *Computational explorationsin cognitive neuroscience: Understanding the mind by simulating the brain.* MA: The MIT Press.

Palmer, C. & Hutchins, S. (2006). What is musical prosody? *Psychology of Learning and Motivation*, *46*, 245–278.

Papoušek, M., Bornstein, M. H., Nuzzo, C., Papoušek, H., & Symmes, D. (1990). Infant responses to prototypical melodic contours in parental speech. *Infant Behaviour and Development*, *13*, 539–545.

Papoušek, M. & Papoušek, H. (1991). The meaning of melodies in motherese in tone and stress languages. infant behaviour and development , 14, 415-440. *Infant Behaviour and Development*, *14*, 415–440.

Papoušek, M., Papoušek, H., & Haekel, M. (1987). Didactic adjustments in fathers and mothers speech to their 3-month old infants. *Journal of Psycholinguistic Research*, *16*(5), 492–516.

Patel, A. (2006). An empirical method for comparing pitch patterns in spoken and musical melodies: A comment on j.g.s pearl's eavesdropping with a master: Leos janácek and the music of speech.. *Empirical Musicology Review*, *1*(3), 166–169.

Patel, A. (2007). *Language and Music as Cognitive Systems*, chap. Language, music, and the brain: A resource-sharing framework, pp. 204–223. Oxford: Oxford University Press.

Patel, A., Peretz, I., Tramo, M., & Labreque, R. (1998). Processing prosodic and musical patterns: A neuropsychological investigation. *Brain and language*, *61*(1), 123–144.

Patel, A. D. (2008). *Music, Language and the Brain.* New York: Oxford University Press.

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of british english and french. *The Journal of the Acoustical Society of America*, *119*(5), 3034–3047.

Pearce, M. & Rohrmeier, M. (2012). Music cognition and the cognitive sciences. *Topics in Cognitive Science*, *4*(4), 468484.

Pellegrino, F. & Andre-Obrecht, R. (2000). Automatic language identification: an alternetive approach to phonetic modelling. *Signal Processing*, *80*, 1231–1244.

Peretz, I. (2006). The nature of music from a biological perspective. *Cognition*, *100*(1), 1–32.

Peretz, I., Coltheart, M. et al. (2003). Modularity of music processing. *Nature neuroscience*, *6*(7), 688–691.

Peretz, I. & Morais, J. (1989). Music and modularity. *Contemporary Music Review*, *4*(1), 279–293.

Peretz, I. & Zatorre, R. (2005). Brain organization for music processing. *Annual Review of Psychology*, *56*, 89–114.

Piaget, J. (1953). *The origins of intelligence.* New York: Routledge.

Piaget, J. (1971). *Measurement and Piaget.*, chap. he theory of stages in cognitive development. , (1971). „ ix, 283 pp., pp. 1–11. New York, NY, US: McGraw-Hill.

Pike, K. (1945). *The intonation of American English.* Ann Arbor: University of Michigan Press.

Pinker, S. (1997). *How the mind works (1st ed.).* New York: Norton.

Platt, J. (1998). *Advances in Kernel Methods - Support Vector Learning*, chap. Machines using Sequential Minimal Optimization. Cambridge. MA: MIT Press.

Plunkett, K. & Elman, J. L. (1997). *Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations*, chap. The methodology of simulations, pp. 19–30. TheMIT press.

Plunkett, K. & Sinha, C. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology*, *10*(3), 209–254.

Port, R. (2003). Meter and speech. *Journal of Phonetics*, *31*(3-4), 599–611.

Povel, D. & Essens, P. (1985). Perception of temporal patterns. *Music Perception*, *2*, 411–440.

Purwins, H., Grachten, M., Herrera, P., Hazan, A., Marxer, R., & Serra, X. (2008). Computational models of music perception and cognition ii: Domain-specific music processing. *Physics of Life Reviews*, *5*(3), 169 – 182.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Quinn, P. & Eimas, P. (2000). The emergence of category representations during infancy: Are separate perceptual and conceptual processes required? *Journal of Cognition and Development*, *1*(1), 55–61.

Quinn, P., Eimas, P., & Rosenkrantz, S. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, *22*, 463–475.

Quinn, P. C. (2007). *Blackwell Handbook of Childhood Cognitive Development*, chap. Early Categorization: A New Synthesis, pp. 84–101. Blackwell Publishers Ltd, Malden, MA, USA.

Quinn, P. C. & Johnson, M. (1997). The emergence of perceptual category representations in young infants: a connectionist analysis. *Journal of Experimental Child Psychology*, *66*(2), 236–263.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition,*, *73*(3), 265–292.

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of experimental psychology: general*, *118*(3), 219.

Ringeval, F. & Chetouani, M. (2008). Exploiting a vowel based approach for acted emotion recognition. In A. Esposito, N. Bourbakis, N. Avouris, & I. Hatzilygeroudis (Eds.) *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, *Lecture Notes in Computer Science*, vol. 5042, pp. 243–254. Springer Berlin / Heidelberg.

Rock, A. M., Trainor, L. J., & Addison, T. L. (1999). Distinctive messages in infant-directed lullabies and play songs. *Developmental Psychology*, *35*(2), 527–534.

Roederer, J. (1984). The search for a survival value of music. *Music Perception*, *1*, 350–356.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192–233.

Rouas, J. L., Farinas, J., Pellegrino, F., & Andrbrecht, R. (2005). Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication*, *47*, 436–456.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, chap. A General framework for parallel distributed processing., pp. 45–76. MA: The MIT Press.

Saffran, J. & Griepentrog, G. (2001). Absolute pitch in infant auditory learning: evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74–85.

Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*, 606–621.

Saffran, J., Reeck, K., Niebuhr, A., & Wilson, D. (2005). Changing the tune: the structure of the input affects infants? use of absolute and relative pitch. *Developmental Science*, *8*(1), 1–7.

Salselas, I., Hazan, A., Herrera, P., & Purwins, H. (2008). The development of melodic representations at early age: Towards a computational model. Tech. rep., EmCAP Project.

Salselas, I. & Herrera, P. (2010). Music and speech in early development: automatic analysis and classification of prosodic features from two portuguese variants. *Journal of Portuguese Linguistics*, *9*(2), 11–35.

Santen, J. P. & Olive, J. (1989). The analysis of contextual effects on segmental duration. *Computer, Speech and Language*, *4*, 359–390.

Santos, R. S. (2005). Banco de dados para projecto de aquisição do ritmo. universidade de são paulo, fflch - departamento de lingüística.

Schafer, G. & Mareschal, D. (2001). Modeling infant speech sound discrimination using simple associative networks. *Infancy*, *2*(1), 7–28.

Schön, D., Boyer, M., Moreno, S., Besson, M., Peretz, I., & Kolinsky, R. (2008). Songs as an aid for language acquisition. *Cognition*, *106*(2), 975–983.

Schoner, G. (2008). *The cambridge handbook of computational psychology*, chap. Dynamical systems approaches to cognition, pp. 101–126. Cambridge University Press.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.

Sergeant, D. & Roche, S. (1973). Perceptual shifts in the auditory information processing of young children. *Psychology of Music*, *1*(39), 39–48.

Shrager, J. & Siegler, R. (1998). Scads: A model of children's strategy choices and strategy discoveries. *Psychological Science*, *9*(5), 405–410.

Shultz, T., Schmidt, W., Buckingham, D., & Mareschal, D. (1995). *Developing cognitive competence: New approaches to process modeling*, chap. Modeling cognitive development with a generative connectionist algorithm, pp. 205–261. Erlbaum, Hillsdale, NJ.

Shultz, T. R. (2003). *Computational Developmental Psychology.* The MIT Press.

Shultz, T. R. & Marechal, D. (1997). Rethinking innateness, learning, and constructivism: Connectionist perspectives on development. *Cognitive Development*, *12*, 563–586.

Shultz, T. R. & Sirois, S. (2008). *The cambridge handbook of computational psychology*, chap. Computational models of developmental psychology, pp. 451–476. Cambridge University Press.

Siegler, R. (1989). Mechanisms of cognitive development. *Annual Review of Psychology*, *40*(1), 353–379.

Siegler, R. (1996). *The five to seven year shift: The age of reason and responsibility*, chap. Unidimensional thinking, multidimensional thinking, and characteristic tendencies of thought, pp. 63–84. Chicago: University of Chicago Press.

Skinner, B. F. (1969). *Contingencies of reinforcement*. New York: Appleton-Century-Croft.

Slaney, M. & McRoberts, G. (2003). Babyears: A recognition system for affective vocalizations. *Speech Communication*, *39*, 367–384.

Slater, A. (1995). Visual perception and memory at birth. *Advances in infancy research*, *9*, 107–162.

Slevc, L., Rosenberg, J., & Patel, A. (2009). Making psycholinguistics musical: Self-paced reading time evidence for shared processing of linguistic and musical syntax. *Psychonomic bulletin & review*, *16*(2), 374–381.

Sloutsky, V. (2010). From perceptual categories to concepts: What develops? *Cognitive science*, *34*(7), 1244–1286.

Soley, G. & Hannon, E. (2010). Infants prefer the musical meter of their own culture: a cross-cultural comparison. *Developmental psychology*, *46*(1), 286–292.

Steedman, M. (1977). The perception of musical rhythm and metre. *Perception*, *6*(5), 555–570.

Sternberg, R. J. (1984). *Mechanisms of cognitive development*. New York: Freeman.

Sun, R. (2008). *The Cambridge Handbook of Computational Cognitive Modeling*, chap. Introduction to computational cognitive modeling, pp. 3–19. Cambridge University Press.

Sun, R., Coward, L., & Zenzen, M. (2005). On levels of cognitive modeling. *Philosophical Psychology*, *18*(5), 613–637.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9–44.

Temperley, D. (2004). *The cognition of basic musical structures*. MA: The MIT Press.

Tenney, J. & Polansky, L. (1980). Temporal gestalt perception in music. *Journal of Music Theory*, *24*(2), 205–241.

Thelen, E. & Smith, L. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge MA: The MIT Press.

Thiessen, E., Hill, E., & Saffran, J. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71.

Thomas, D. A. (1995). *Music and the origins of language*. Cambridge: Cambridge University Press.

Thomas, M. & McClelland, J. (2008). *The Cambridge handbook of computational psychology*, chap. Connectionist models of cognition, pp. 23–58. Cambridge University Press.

Thompson, N. C., Cranford, J. L., & Hoyer, E. (1999). Brief-tone frequency discrimination by children. *Journal of Speech, Language and Hearing Research*, *42*, 1061–1068.

Thorpe, L. A. & Trehub, S. E. (1989). Duration illusion and auditory grouping in infancy. *Developmental Psychology*, *25*(1), 122–127.

Todd, N. (1994). The auditory primal sketch: A multiscale model of rhythmic grouping. *Journal of new music Research*, *23*(1), 25–70.

Trainor, L. (2005). Are there critical periods for musical development? *Developmental Psychobiology*, *46*(3), 262–278.

Trainor, L. & Desjardins, R. (2002). Pitch characteristics of infant-directed speech affect infants ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*(2), 335–340.

Trainor, L. J., Clark, E. D., Huntley, A., & Adams, B. A. (1997). The acoustic basis of preferences for infant-directed singing. *Infant Behaviour and Development*, *20*(3), 383–396.

Trainor, L. J., Wu, L., & Tsang, C. D. (2004). Long-term memory for music: infants remember tempo and timbre. *Developmental Science*, *7*(3), 289–296.

Trehub, S. (2000). *The origins of music*, chap. Human processing predispositions and musical universals, pp. 427–448. Cambridge, MA: The MIT Press.

Trehub, S. (2001). Musical predispositions in infancy. *Annals of the New York Academy of Sciences*, *930*, 1–16.

Trehub, S. (2003). The developmental origins of musicality. *Nature neuroscience*, *6*(7), 669–673.

Trehub, S. & Hannon, E. (2006). Infant music perception: Domain-general or domain-specific mechanisms? *Cognition*, *100*(1), 73–99.

Trehub, S. & Trainor, L. (1998). *Advances in infancy research*, chap. Singing to infants: Lullabies and play songs, pp. 43–78. Ablex Publishing Corporation.

Trehub, S. E., Unyk, A. M., & Trainor, L. J. (1993). Maternal singing in cross-cultural perspective. *Infant behavior & development*, *16*(3), 285–295.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

van Ooyen, B., Bertoncini, J.and Sansavini, A., & Mehler, J. (1997). Do weak syllables count for newborns? *Journal of the Acoutic Sociaty of America*, *102*, 3735–3741.

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

Warren, R. (1974). Auditory temporal discrimination by untrained listeners. *Cognitive Psychology*, *15*, 495–500.

Warren, R. M. (1993). *Thinking in Sound: the cognitive psychology of human audition*, chap. Perception of acoustic sequences: global integration versus temporal resolution., pp. 37–68. Oxford Science Publications.

Weinberger, N. (2004). Specific long-term memory traces in primary auditory cortex. *Nature Reviews Neuroscience*, *5*(4), 279–290.

Werker, J. & McLeod, P. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology / Revue canadienne de psychologie*, *43*(2), 230–246.

Werner, L. A. & VandenBos, G. R. (1993). Developmental psychoacoustics: What infants and children hear. *Hospital and Community Psychiatry*, *44*, 624–626.

Westermann, G., Sirois, S., Shultz, T., & Mareschal, D. (2006). Modeling developmental cognitive neuroscience. *Trends in Cognitive Sciences*, *10*(5), 227–232.

Wibke, G., Ulrike, L., & Thomas, O. (2010). Effects of music therapy in the treatment of children with delayed speech development-results of a pilot study. *BMC Complementary and Alternative Medicine*, *10*, 1–39.

Witten, I. H. & Frank, E. (2005). *Data mining: practical machine learning tools and techniques, Second Edition.* San Francisco: Elsevier.

Woodrow, H. (1909). *A quantitative study of rhythm:The effect of variations in intensity, rate and duration.* New York : The Science press.

Yoshida, K. A., Iversen, J. R., Patel, A. D., Mazuka, R., Nito, H., Gervain, J., & Werker, J. F. (2010). The development of perceptual grouping biases in infancy: A japanese-english cross-linguistic study. *Cognition*, *115*(2), 356–361.

Younger, B. & Gotlieb, S. (1988). Development of categorization skills: Changes in the nature or structure of infant form categories? *Developmental Psychology*, *24*(5), 611–619.

Zatorre, R., Belin, P., & Penhune, V. (2002). Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, *6*(1), 37–46.