Music and speech in early development: automatic analysis and classification of prosodic features from two Portuguese variants

INÊS SALSELAS
PERFECTO HERRERA

Abstract

In the present study we aim to capture rhythmic and melodic patterning in 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. Machine learning experiments were conducted in order to automatically discriminate between language variants for speech, vocal songs and between interaction contexts. Descriptors related with rhythm exhibited strong predictive ability for both speech and singing language variants' discrimination tasks, presenting different rhythmic patterning for each variant. 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. With respect to discriminating interaction contexts, 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. These prosodic differences may lead to further research on their influence in the development of the infant's musical representations.

1. Introduction

1.1. Aims

Early experience has a fundamental role in brain development. During infancy, the brain is developing 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 it (Patel, 2008). Exposure to the information to which infants are subjected, along with genetic influence, is determinant to the strengthening of neural communication paths, synaptic formation and organization. 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 (Papousek & Papousek, 1991; Trehub, Unyk, & Trainor, 1993). Regarding singing, caregivers typically 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, Clark, Huntley, & Adams, 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, Trainor, & Addison, 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 (Papousek, Papousek, & Haekel, 1987). In addition to engaging and maintaining the infant's attention, these distinctive modifications play an important role for indicating different communicative intentions to preverbal 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, Bornstein, Nuzzo, Papoušek, & Symmes, 1990). These prototypical shapes have been considered cross--linguistic universals (Papousek & Papousek, 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 passively 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 preverbal 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 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 3, different classification experiments will be reported. Section 4 presents the discussion of the results obtained, and the last section presents our conclusions.

1.2. Background

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. Both musical and speech prosody 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 the 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, Nespor, & Mehler, 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 & Vigário, 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, Wu, & Tsang, 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 400Hz, which suggests that they probably rely on distinctions between vowels and consonants to accomplish the discrimination task (Mehler, Dupoux, Nazzi, & Dehaene-Lambertz, 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, Farinas, Pellegrino, & André-Obrecht, 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, Garcia, Duarte and Galves (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 and 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, Cohn, & Moore, 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.

2. Methods

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: (i) affection, a positive affect to provide comfort to the infant such as "Ohhh my sweet baby"; (ii) disapproval, a negative affect such as "No!! Don't do that!"; (iii) questioning, a more complex sound sequence such as "Would you like to have a cookie?"; and (iv) 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 I.

Table I: 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 several interaction contexts. Therefore, for its variety in content, the corpus can be considered representative.

2.2. Discrimination system model

2.2.1. 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 1 illustrates a pitch contour stylisation from Prosogram.

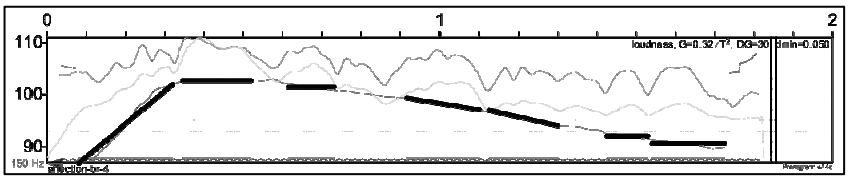

Figure 1: Illustration of the Prosogram of an affection instance ("hmmmm nham nham nham nham nham"). The horizontal axis represents time in seconds and the vertical axis shows semitones (relative to 1 Hz). Intensity is represented in the green line, fundamental frequency in the blue line, and the intensity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

Prosogram is a suitable tool for studying music and language (Patel, Iversen, & Rosenberg, 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 200Hz. 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, Farinas, Pellegrino, & André-Obrecht, 2005; Ringeval & Chetouani, 2008) was used to evaluate Prosogram, as well as vowel onset and offset detection. VER is defined follows:

$$VER = 100.\ [(N_{del} + N_{ins})/N_{vow}]\ \ \% \qquad (1)$$

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 II 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 and Chetouani (2008).

Table II: 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.

2.2.2. 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, Bertoncini, Sansavini, & Mehler, 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, Floccia, Nazzi, & Mehler, 1995).

2.2.3. 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, Grabe, & Nolan, 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.

2.2.4. 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.

2.2.5. 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.

[1] F measure = (2 * recall*precision) / (recall+precision)

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.

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 & Vigário, 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 III 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.

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 III: 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)

Table IV presents the mean, standard deviation and p-value for rhythm--related descriptors shown to be relevant in language discrimination tasks (see sub-section 2.2.2 *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.

Table IV: 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

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.

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, Wu, & Tsang, 2004; Mehler, Dupoux, Nazzi, & Dehaene-Lambertz, 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 2.2 *Discrimination system model*). Statistical information of the utterances in dataset built for this experiment is provided in Table V. Once again, speech rate is higher for Brazilian Portuguese, as had occurred with for speech.

Table V: Basic statistical information about the utterances grouped 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)

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:

- 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 VI 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.

Table VI: 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

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 3.1. *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).

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 VII, 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.

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 VIII 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.

Table VII: 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

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.

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 (Papousek & Papousek, 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 IX), and European Portuguese (EP) Affection, Disapproval, Question (A-EP, D-EP, Q-EP, respectively, in Table IX). The distribution of instances per classes as well as the corresponding statistical information is shown in Table IX. 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 VIII: 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

Table IX: 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 X, that shows the confusion matrix, communicative contexts are confused across variants.

Table X: 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

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 IV). As referred before, the two Portuguese variants are considered to have distinct rhythm patterning (Frota & Vigário, 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 (Papousek, Papousek, & Haekel, 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 VI), 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, Iversen, & Rosenberg, 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 (Papousek & Papousek, 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 X), 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 teh 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, Dupoux, Nazzi, & Dehaene-Lambertz, 1996; Nazzi & Ramus, 2003; Ramus, Nespor, & Mehler, 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. Conclusions

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 IV). 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 discriminatios 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 preverbal 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 future work we will aim to build a developmental model exploring the fact that prosodic features present in infant-directed speech and singing may affect the infant's development of melodic representations.

Acknowledgments

We thank all the caregivers who were willing to collaborate in this study, Raquel Santos for providing audio material and Piotr Holonowicz for performing the vowel onset evaluation. The Music Technology Group, Universitat Pompeu Fabra, supported this work. Inês Salselas was supported by a grant from Barcelona Media.

References

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.

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.

Diamond, A. & Goldman-Rakic, P. (1989) Comparison of human infants and rhesus monkeys on Piaget's AB task: evidence for dependence on dorsolateral prefrontal cortex, *Experimental Brain Research,* **74** (1), 24-40.

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

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

Frota, S. & Vigário, M. (2001) On the correlates of rhythmic discriminations: The european/brazilian Portuguese case, *Probus,* **13**, 247-275.

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

Grabe, E. & Low, E. L. (2002) Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warder (Eds.), *Laboratory Phonology 7* (pp. 515-546). Berlin, Germany: Mouton de Gruyter.

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

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.

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.

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. S., Saint-Georges, C., Laznik, M.--C. et al. (2009) Automatic motherese detection for face-to-face interaction analysis. In *Multimodal Signals: Cognitive and Algorithmic Issues* (pp. 248-255). Berlin / Heidelberg: Springer.

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

Mehler, J., Dupoux, E., Nazzi, T. & Dehaene-Lambertz, G. (1996) Coping with linguistic diversity: The infant's viewpoint. In J. L. Morgan, & K. e. Demuth, *Signal to syntax: bootstrapping from speech to grammar in early acquisition.* Mahwah NJ: Lawrence Erlbaum Associates.

Mertens, P. (2004) The prosogram: semi-automatic transcription of prosody based on a tonal perception model, *Proceedings of Speech prosody 2004*, (págs. 549-552).

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

Nooteboom, S. (1997) The prosody of speech: melody and rhythm. In J. Hardcastle, & J. E. Laver, *The handbook of phonetic sciences* (págs. 640–673). Cambridge, MA: Blackwell.

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

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

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.

Papousek, M., Papousek, 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. D. (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 mpeech", *Empirical Musicology Review*, 166-169.

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, *Journal of Acoustic Society of America*, **119** (5), 3034-3047.

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

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

Platt, J. (1998) Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola (Eds.), *Advances in Kernel Methods – Support Vector Learning,* Cambridge. MA: MIT Press.

Ramus, F. & Mehler, J. (1999) Language identification with suprasegmental cues: a study based on speech resynthesis, *The Journal of the Acoustical Society of America,* **105** (1), 512-21.

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

Ringeval, F. & Chetouani, M. (2008) Exploiting a vowel based approach for acted emotion recognition. In *Verbal and nonverbal features of human-human and human-machine interaction* (Vol. 5042/2008, 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.

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

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.

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

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. E., Unyk, A. M. & Trainor, L. J. (1993) Maternal singing in cross-cultural perspective, *Infant behavior & development,* **16** (3), 285-295.

van Ooyen, B., Bertoncini, J., Sansavini, A. & Mehler, J. (1997) Do weak syllables count for newborns?, *Journal of the Acoustic Society of America,* **102**, 3735-3741.

Vapkin, V. (1982) *Estimation of dependencies based on empirical data.* Verlag: Springer.

Witten, I. H. & Frank, E. (2005) *Data mining: practical machine learning tools and techniques*, Second Edition. San Francisco: Elsevier.

Inês Salselas
Music Technology Group,
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona
Spain
Ines.salselas@upf.edu

Perfecto Herrera
Music Technology Group,
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona
Spain
perfecto.herrera@upf.edu