Cooing, Crying, and Babbling: A Link between Music and Prelinguistic Communication

Michael Byrd, Casady Bowman, and Takashi Yamauchi
(mybrd@neo.tamu.edu, casadyb@neo.tamu.edu, takashi-yamauchi@tamu.edu)
Department of Psychology, Mail Stop 4235
Texas A&M University, College Station, TX 77843 USA

Abstract

Like language, the human capacity to create music is one of the most salient and unique markers that differentiates humans from other species (Cross, 2005). In the following study, the authors show that people's ability to perceive emotions in infants' vocalizations (e.g., cooing and babbling) is linked to the ability to perceive timbres of musical instruments. In one experiment, 180 "synthetic baby sounds" were created by rearranging spectral frequencies of cooing, babbling, crying, and laughing made by 6 to 9-month-old infants. Undergraduate participants (N=145) listened to each sound one at a time and rated the emotional quality of the "synthetic baby sounds." The results of the experiment showed that five acoustic components of musical timbre (e.g., *roll off, mel-frequency cepstral coefficient, attack time and attack slope*) could account for nearly 50% of the variation of the emotion ratings made by undergraduate students. The results suggest that the same mental processes are probably applied for the perception of musical timbres and that of infants' prelinguistic vocalization.

Keywords: Emotion; Language; Music

Introduction

Infants use a variety of vocal sounds, such as cooing, babbling, crying, and laughing, to express their emotions. Infants' prelinguistic vocal communications are highly affective in the sense that they evoke specific emotions—happiness, frustration, anger, hunger, and/or joy—without conveying concrete ideas. In this sense, infants' vocal communication parallels music. Music is highly affective; yet it is conceptually limited (Cross, 2005; Ross, 2009).

The interaction between music and language has attracted much attention recently (Chen-Haffteck, 2011; Cross, 2001; Masataka, 2007). However, despite their similarities, little attention has been paid to the relationship between music and prelinguistic vocalizations (Chen-Haffteck, 2011; Cross, 2001; He, Hotson, & Trainor, 2007; Masataka, 2007). If music and language are highly related, what is the relationship between infants' vocal communications such as babbling, and music?

In the study described below, we analyze acoustic cues of infants' vocalization and demonstrate that emotions created by prelinguistic vocalization can be explained to a large extent by the acoustic cues of sound that differentiate timbres of musical instruments, potentially implicating that the same mental processes are applied for the perception of musical timbres and that of infants' vocalizations.

The paper is organized as follows: we review related work examining the link between prelinguistic vocalization and music followed by an overview of the experiment. After discussing our timbre extraction and sound creation method, we introduce one experiment that investigates the connection between music and prelinguistic communication.

Related Work

Infants begin life with the ability to make different sounds—first cooing and crying, then babbling. Next they form one word, and then two, followed by full sentences and speech. In the first ten months, infants progress from simple sounds that are not expressed in the phonetic alphabet, to babbling, which is an important step in infants learning how to speak (Gros-Louis, West, Goldstein, & King, 2006; Oller, 2000).

Musical instruments and infants' vocalizations both elicit emotional responses, while conveying little information on what the sender is trying to express. Music can have a very powerful effect on its listeners, as we all have a piece of music that will bring back emotions. Music can convey at least three universal emotions, happiness, sadness and fear (Fritz et al., 2009). These emotions are similar to the emotions expressed by infants with their limited sounds (Dessureau, Kurowski, & Thompson, 1998; Zeifman, 2001; Zeskind & Marshall, 1998). Both infants and music convey meaning without the use of words. Infants rely on their voices and non-verbal/non-word sounds to communicate and it is these sounds that inform the listener of how important and of what type of danger the infant is facing, such as being too cold, hungry or of being left alone (Dessureau et al., 1998; Zeifman 2001; Zeskind & Marshall, 1998).

Across cultures, songs sung while playing with babies are fast, high in pitch, and contain exaggerated rhythmic accents, whereas lullabies are lower, slower and softer. Infants will use cues in both music and language to learn the rules of a culture. Motherese, a form of speech used by adults in interacting with infants, often consists of singing to infants using a musical, sing-song voice, that mimics babies' cooing by using a higher pitch. An infant's caregiver will use higher pitch when speaking to an infant, as it helps the infant learn and also draws their attention (Fernald 1989).

In summary, research shows that there is a close link between infants' vocal communication and music. This link is demonstrated through the babbling and cooing sounds used by infants' to communicate, and also by mothers' use of motherese to assist infant's learning of language in a sing-song manner. Infants are able to use the same cues

from both music and language to facilitate learning in both domains. Given these close connections, it is likely that the same mental processes are involved for the perception of instrumental sounds and the perception of infants' vocalizations. The beginning stages of this idea are investigated in one experiment by examining the emotion perception of synthetic baby sounds.

Overview of the Study

In the Emotion Rating Experiment described below, we tested the general hypothesis that the same mental process is involved for the perception of infants' vocalization and that of timbres of musical instruments. More specifically, we hypothesize that the acoustic components of timbre will be significant predictors of emotion. If this is true, then there should be a plausible link between musical timbre and prelinguistic vocal timbre, also indicating a link for mental processing in the two domains. We employed an audio synthesizer program and created 180 different "synthetic baby sounds" by combining spectral frequencies of real baby sounds. In the experiment, our undergraduate participants (N=145) listened to the "synthetic baby sounds" one at a time and rated affective qualities of these sounds. Later, we extracted "musical timbres" from the synthetic baby sounds, and examined the extent to which the emotional ratings made by our undergraduate students were accounted for by the timbres of the synthetic baby sounds.

Timbre is an important perceptual feature of both music and speech. Timbre is defined as the "acoustic property that distinguishes two sounds"—for example, those of the flute and the piano—"of identical pitch, duration, and intensity" (Hailstone et al., 2009; McAdams & Cunible, 1992). The classic definition of timbre states that two different timbres result from the sound of different amplitudes (of harmonic components) of a complex tone in a steady state" (Helmholtz, 1885). Timbre is a sound quality that encompasses the aspect of a sound that is used to distinguish it from other sounds of the same pitch, duration, and loudness.

The timbre properties of *attack time, attack slope, zero-cross, roll off, brightness, mel-frequency cepstral coefficients, roughness, and irregularity* are well known in music perception research as the main acoustic cues that correlate with the perception of timbre of musical instruments (Hailstone et al., 2009). Our assumption is that if infants' vocal sounds are perceived in the same manner as the timbres of musical instruments are perceived, these same acoustic properties can account for the perception of emotions in infants' vocalization.

Using principal components analysis (PCA), we summarized emotional ratings made by our undergraduate participants into two principal dimensions, to reduce the data, and applied a stepwise regression to evaluate the extent to which our predictors—the acoustic timbre components—accounted for emotion ratings for synthesized baby sounds.

Below, we briefly describe our timbre extraction method and the method of creating "synthetic baby sounds."

Timbre Extraction

This section describes acoustic cues relating to timbre in detail, as well as the computational procedure of extracting these cues. The purpose of using these acoustic cues is to act as predictors in regression analyses that can explain perceived emotions of our "synthetic baby sounds." The acoustic cues were chosen based on their use in musical timbre (see Lartillot & Toiviainen, 2007).

Eight acoustic properties of timbre: attack time, attack slope, zero-cross, roll off, brightness, mel-frequency cepstral coefficients, roughness, and irregularity were extracted from all sound stimuli using MIRToolbox in Matlab (Lartillot, Toiviainen, & Eerola, 2008). These acoustic properties are known to contribute to the perception of timbre in music independently of melody and other musical cues (Hailstone et al., 2009). The acoustic features were extracted from synthesized sounds rated in the Emotion Rating Experiment.

Attack time is the time in seconds it takes for a sound to travel from amplitude of zero, to the maximum amplitude of a given sound signal, or more simply the temporal duration. Some features of timbre such as attack time contribute to the perception of emotion in music (Gabrielsson & Juslin, 1996; Juslin, 2000; Loughran, Walker, O'Neill & O'Farrell, 2001); which suggests that features of timbre can at least in part determine the emotional content of music (Hailstone et al., 2009).

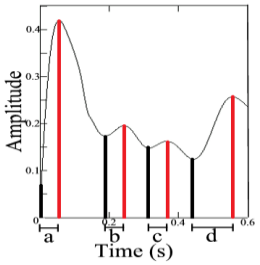

Figure 1. Attack times of an audio file. *A* through *d* are separate attack times; indicateb by the distance from the black line, to the the red line.

Attack time is computed using the equation of a line, y = mx + b, it is part of a sounds amplitude envelope where *m* is the slope of the line and *b* is the point where the line crosses the vertical axis (*t*=0). Figure 1 gives an illustration of attack time. The horizontal segments below the x-axis indicate the time it takes in seconds to achieve the maximum peak of each frame for which the attack time was calculated.

Attack slope is the attack phase of the amplitude envelope of a sound, also interpreted as the average slope leading to the attack time. This can also be calculated using the equation of a line y = mx +b, where *m* is the slope of the line and *b* is the point where the line crosses the vertical axis (*t*=0), see Figure 2. The red line in Figure 2 indicates the slope of the attack.

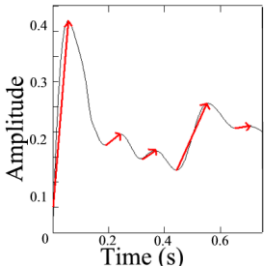

Figure 2. Attack slope of a audio file. The red arrow indicates the duration (attack time) for which the attack slope is calculated.

Zero-cross is the number of times a sound signal crosses the x-axis, this accounts for noisiness in a signal and is calculated using the following equation where sign is 1 for positive arguments and 0 for negative arguments. X[n] is the time domain signal for frame t.

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} | \, sign(x[n]) - sign(s[n-]) $$

Roll off is the amount of high frequencies in a signal, which is specified by a cut-off point. The roll-off frequency is defined as the frequency where response is reduced by -3 dB. This is calculated using the following equation where Mt is the magnitude of the Fourier transform at frame *t* and frequency bin *n*. Rt is the cutoff frequency, see Figure 3.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{R_t} M_t[n]$$

Brightness is the amount of energy above a specified frequency, typically set at 1500 Hz – this is related to spectral centroid. The term brightness is also used in discussions of sound timbres, in a rough analogy with visual brightness. Timbre researchers consider brightness to be one of the strongest perceptual distinctions between sounds. Acoustically it is an indication of the amount of high-frequency content in a sound, and uses a measure such as the spectral centroid, see Figure 3.

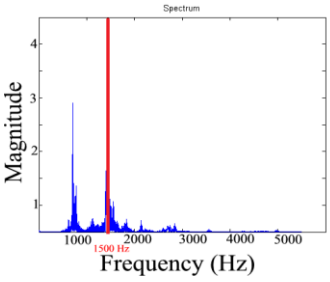

Figure 3. Brightness of an audio file. To the right of the red dashed line is the amount of energy above 1500 Hz, or the brightness of the sound.

Roughness is sensory dissonance, the perceived harshness of a sound; this is the opposite of consonance (harmony) within music or even a single tone harmonics. Both consonance and dissonance are relevant to emotion perception (Koelsch, 2005). Roughness is calculated by computing the peaks within a sound's spectrum and measuring the distance between peaks, dissonant sounds have irregularly placed spectral peaks as compared to consonant sounds with evenly spaced spectral peaks.

Formally, roughness is calculated using the following equation where a_j and a_k are the amplitudes of the components, and g (f_{cb}) is a 'standard curve.' This was first proposed by (Plomp & Levelt, 1965).

$$\rho = \frac{\sum_{j,k}^{n} a_j \cdot a_k \cdot g(f_{cb})}{\sum_{j}^{n} a_j^2}$$

Following extraction of the value for roughness from the sound stimuli, principal components analysis was used to reduce the dimensions of the roughness data.

Mel-frequency Cepstral Coefficients (mfcc) represent the power spectrum of a sound. This power spectrum is based on a linear transformation from actual frequency to the Mel-scale of frequency. The Mel scale is based on a mapping between actual frequency and perceived pitch as the human auditory system does not perceive pitch in a linear manner. Mel-frequency cepstral coefficients are the dominant features used in speech recognition as well as some music modeling (Logan, 2001). Frequencies in the Mel scale are equally spaced, and approximate the human auditory system more closely than a linearly spaced frequency bands used in a normal cepstrum. Due to large data output, prior to analyses mfcc data were reduced using principal components analyses to create a workable set of data. A cutoff criterion of 80% was used to represent the variability in the original mfcc data. Figure 4 shows the numerical Mel-frequency cepstral coefficient rank values for the 13 mfcc components returned. Thirteen components are returned due to the concentration of the signal information in only a few low-frequency components.

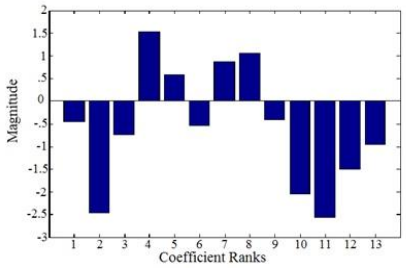

Figure 4. Mel-frequency cepstral coefficients (mfcc) of an audio file. This figure shows the acoustic component mfcc. Each bar represents the numerical (rank coefficient) value computed for the thirteen components returned.

Irregularity of a spectrum is the degree of variation between peaks of a spectrum (Lartillot et al., 2008). This is calculated using the following equation where irregularity is the sum of the square of the difference in amplitude between adjoining partials in a sound.

$$\frac{\sum_{k=1}^{N}(a_k - a_{k+1})^2}{\sum_{k=1}^{N} a_k^2}$$

Creating Synthetic Baby Sounds

We created 180 short, 2 second, synthetic baby sounds from ten real infant sounds: five males and five females ranging from ages 6 to 9 months making screaming, laughing, crying, cooing and babbling sounds. These sounds were chosen to create novel stimuli emulating human prelinguistic sounds. Among these sounds, four (one screaming boy, one crying boy, one screaming girl and one crying girl) were audio-recorded directly from two volunteer infants in Nacogdoches, Texas using an Olympic Digital Voice WS-400S recorder. The sounds of babbling and cooing boys and girls were taken from audio-files downloaded from a sound effects website (http://www.freesounds.org), and the sounds of laughing boy and girl were taken from files downloaded from YouTube (http://www.youtube.com).

These infant sounds were decomposed by four laboratory assistants into amplitude and spectral frequency components by applying fast Fourier transform using a sound editing software program (SPEAR, Klingbeil, 2005). Arbitrarily chosen spectral frequencies of one sound (e.g., a babbling sound of a boy) were mixed with arbitrarily chosen spectral frequencies of another sound (e.g., cooing girl) and then modified by means of amplitude, or shifting frequencies, to convey one of the basic emotions, happy, sad, anger, or fear (Ekman, 2002).

For each sound pair, four sounds were created to sound happy, sad, angry, and fearful. In this manner, each sound pair (45 pairs in total, all possible pairs of the 10 real sounds), was used to create four affective sounds, which was decided subjectively by the laboratory assistants. The total 180 sound stimuli were normalized and white noise was taken out prior to and after creation of each sound stimulus.

Emotion Rating Experiment

The goal of the experiment was to obtain empirical ratings of college students examining the emotional quality of the synthetic baby sounds that we created. To analyze the link between emotion ratings and acoustic cues, a stepwise regression analysis was employed.

Participants. A total of 145 undergraduate students (73 males, 73 females) participated in this experiment for course credit. Participants were randomly assigned to one of two groups that listened to 90 or 89 sounds of 179 total sounds**.** Stimuli were randomly assigned to one of two groups; no participants were in both groups.

Materials. Stimuli were taken from the 180 synthetic baby sounds that were created from a group of a total of ten recorded real infants' sounds (see the "Creating Synthetic Baby Sounds" section for the details of the sound creation).

Procedure. Participants were presented with 90/89 sounds using customized Visual Basic software through JVC Flats stereo headphones. Each stimulus's maximum volume was adjusted and normalized. Participants were instructed to listen to sound stimuli, and rate each sound on five emotion categories, happy, sad, angry, fearful, and disgusting (Ekman, 1992; Johnson-Laird & Oatley, 1989). Each scale ranged from 1 to 7—1 being *strongly disagree* (the degree to which the stimuli, sounded like one of the five emotions), and 7 being *strongly agree*. Stimuli were presented in a random order.

Results

This section starts with descriptive statistics of emotion ratings followed by the results from stepwise regression analysis, which examined the extent to which emotion ratings given to the synthetic baby sounds were explained by their timbre properties. For the regression analysis, average emotion scores were calculated for individual synthetic sounds by collapsing over individual participants, yielding a 179 sounds x 5 emotion dimension matrix. By applying principal component analysis (PCA), this matrix was summarized in a 179 x 2 matrix with the two columns corresponding to two principal components identified by the PCA procedure. The first two orthogonal components explained 88.1% and 7.1% of the variance of the emotion rating data, respectively.

Descriptive Statistics. Behavioral data, Figure 5, shows overall observations for each emotion from the emotion rating data. From the whiskers of the box plot for the emotion data, it is apparent that there is variation within the data. The highest rating for the emotion data did not exceed a value of 6, on the scale of 1-7. The median of the ratings for emotion varied between approximately 2.5 and 4.75 within the emotion rating data.

For all 179 sounds rated, most were rated as angry, indicated by the median of the data for anger. The sounds were rated least like the emotion happy, as the median for this emotion was the lowest for all sounds rated on the five emotions.

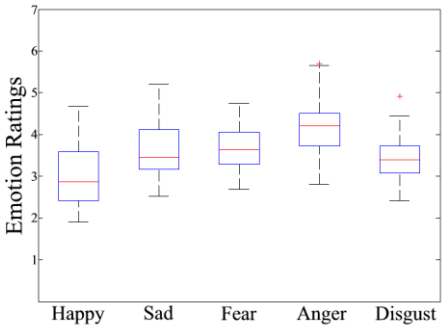

Figure 5. Box plot of observations for emotion ratings. Each box in the figure indicates one emotion rated by participants. The median is indicated by the red line in the center of each box, and the edges indicate the 25th and 75th percentiles, the whiskers of each plot indicate the extreme data points, and outliers are plotted outside of the whiskers.

Regression analysis. A step-wise regression analysis was used to analyze the collected rating data and timbre components, to determine which component could best explain the emotion rating data. Seventeen total predictors were used in the stepwise regression to analyze the emotion ratings made by participants. These were *attack time*, *attack slope*, *zero-cross*, *roll off*, *brightness*, *mel-frequency cepstral coefficients 1-6*, *roughness 1-4*, and *irregularity*. Due to large data output, mfcc data were reduced using principal components analyses to create a workable set of data. There were originally 13 numerical Mel-frequency cepstral coefficient rank values returned. These 13 rank values were reduced to 6, accounting for 78% of the total mfcc data. Roughness was also reduced in the same way using PCA, from 79 components to four components that described 80% of the original roughness data. These predictors were used to analyze the emotion ratings made by participants.

The results of the regression for the first principal component (PCA1) indicated four acoustic features significantly predicted emotion ratings; roll off (β = -.386, p< .001), mfcc 6 (β =. 218, p< .001), attack time (β =. 248, p< .001), and mfcc 3 (β = -.202, p< .002), and attack slope (β = .034, p< .034), see Table 1 for percent explained by principal component 1.

Table 1: Significant acoustic components for emotion PCA 1and PCA 2

Predictors	PCA 1	PCA 2
% explained	88%	7.1%
Attack time	.23***	.31***
Attack slope	.12*	
Irregularity		-.16*
Mfcc 1		-.24**
Mfcc 3	-.19**	
Mfcc 6	.22***	
Roughness		.21**
Zero-cross		.25**
Roll off	-.41***	

* p <.05, ** p < .01, and *** p < .001.

The second principal component (PCA 2) showed that five acoustic features significantly predicted emotion ratings; mfcc 1 (β = -.244, p< .001), zero cross (β = .250, p< .002), attack time (β = .305, p< .000), roughness 2 (β = .208, p< .006), and irregularity (β = -.159, p<. 024), (Table 1).

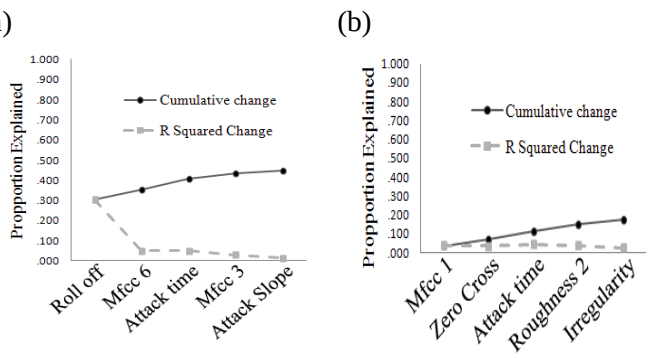

Figure 6. R-squared. Emotion judgment principal components 1 (PCA 1) and 2 (PCA 2). This figure shows the proportion of R-squared contributed for each addition of a predictor to the model for principal component I and II from the emotion judgments.

Figure 6 shows the proportion of R-squared contributed for each addition of a predictor to the model for PCA 1 – (a) and PCA 2 – (b). Looking at the values of R-squared, it is apparent that roll-off was best able to describe the emotion ratings, accounting for 30% of the emotion ratings for PCA 1. The second principal component does show several significant acoustic cues that predict emotion; however, none are as strong as in the first principal component.

General Discussion

Music and language are perhaps two of the most cognitively complex and emotionally expressive sounds invented by humans. Recently, the evolutionary origins of music and language have attracted much attention in researchers of a broad spectrum (Cross, 2001, 2005; Hauser et al., 2002; Kirby, 2007). The present study, examining the relationship between infants' vocalizations—cooing, babbling, crying and screaming—and the perception of musical timbres, suggests that the link between music and language can go even further back to the prelinguistic level of development.

Our Emotion Rating Experiment indicates that nearly 50% of emotions created by synthetically produced infant sounds can be explained by a small number of acoustic cues pertaining to musical timbres. Among those, *roll off*, which quantifies the amount of high frequencies in a signal, turned out to be the most important cue. The second most important property, *mfcc* (*mel-frequency cepstral coefficients)*, corresponds to perceived pitch in the human auditory system, and are the dominant features used in speech recognition and music modeling (Logan, 2001). Given these findings, we conjecture that high-frequency sounds are probably taken as the robust cue of emotion attribution, and more fine-grained distinctions of emotion are made by extracting speech-related cues.

The ability to discriminate sounds is said to be present even in primitive animals such as carp (Chase, 2001), implying that this ability evolved early in history. Some animals have sounds and or calls that can convey the emotions of finding something of interest or of fear (Hauser,

Chomsky, & Fitch, 2002). Such abilities were probably present even before music was fully developed in the current form.

Acknowledgments

We would like to thank Na Yung Yu and Ricardo Gutierrez-Osuna for their valuable comments. The first two authors, MB and CB, contributed to this study an equal amount and the order of their authorship was determined by a coin toss.

References

Chase, A. R. (2001). Music Discriminations by carp (Cyprinus carpio). *Animal Learning and Behaviour, 29*, 336-353.

Cross, I. (2001). Music, mind and evolution. *Psychology of Music, 29*, 95-102.

Cross, I. (2005). Music and meaning, ambiguity and evolution. In D. Miell, R. MacDonald, D. Hargreaves (Ed.), *Musical Communication* (pp. 27-43). New York: Oxford University Press.

Dessureau, B. K., Kurowski, C. O., & Thompson, N. S. (1998). A reassessment of the role of pitch and duration in adults' responses to infant crying. *Infant Behavior and Development, 21*, 367-371.

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*, 550-553.

Fernald, A. (1989). Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message? *Child Development, 60*, 1497-1510.

Fritz, T., Jenschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Koelsch, S. (2009). Universal Recognition of Three Basic Emotions in Music. *Current Biology, 19*, 573-576.

Gabrielsson, A., & Juslin, P. N. (1996). Emotional expression in music performance between the performer's intention and the listener's experience. *Psychology of Music, 24*, 68-91.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development, 30*, 112-119.

Hailstone, J. C., Omar, R., Henley, S., Frost, C., Kenward, M., & Warren, J. D. (2009). It's not what you play, it's how you play it: Timbre affects perception of emotion in music. *Quarterly Journal of Experimental Psychology, 62*, 2141-2155.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science, 22*, 1569-1580.

He, C., Hotson, L., & Trainor, L. J. (2007). Mismatch Responses to Pitch Changes in Early Infancy. *Journal of Cognitive Neuroscience, 19*, 878-892.

Helmholtz, H. v. (2005). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. London: Longmans, Green, and Co.

Johnson-laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition & Emotion, 3*, 81-123.

Juslin, P. N. (2000). Cue utilization in communication of emoting in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1797-1813.

Kirby, S. (2007). The evolution of language. In R. Dunbar & L. Barrett (Eds.), *Oxford handbook of evolutionary psychology* (pp. 669-681). Oxford: Oxford University Press.

Klingbeil, M. (2005). Software for spectral analysis, editing, and synthesis *Proceeding of the ICMC* (pp. 107-110). Barcelona Spain.

Koelsch, S. (2005). Neural substrates of processing syntax and semantic in music. *Current Opinion in Neurobiology, 15*, 207-212.

Lartillot, O., Toiviainen, P., Eerola, T. (2008) A Matlab Toolbox for Music Information Retrieval. In, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (eds.), *Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization*, (pp 261-268). New York: Springer.

Logan, B., & Robinson, T. (2001). Adaptive model-based speech enhancement. *Speech Communication, 34*, 351-368.

Loughran, R., Walker, J., O'Neill, M., O'Farrell, M. (2001). The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification. in Proc. of the 6th International Conference on Music Information Retrieval (ISMIR), 2005, Finland, (pp. 1825–1828).

Masataka, N. (2007). Music, evolution and language. *Developmental Science, 10*, 35-39. McAdams, S., & Cunible, J. C. (1992). Perception of timbral analogies. *Philosophical Transactions: Biological Sciences, 9*, 336-383.

Olivier Lartillot, Petri Toiviainen, "A Matlab Toolbox for Musical Feature Extraction From Audio", *International Conference on Digital Audio Effects*, Bordeaux, 2007.

Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum

Plomp, R., & Levelt, W. J. M. (1965). *Tonal consonance and critical bandwidth*. Soesterberg: Institute for Perception RVO-TNO, National Defense Research Council T.N.O.

Ross, B. (2009). Challenges facing theories of music and language co-evolution. *Journal of the Musical Arts in Africa, 6*, 61-76.

Zeskind, P., S, & Marshall, T., R. (1988). The Relation between Variations in Pitch and Maternal Perceptions of Infant Crying. *Child Development, 59*, 193-196.