Chapter 12
Beyond the Language Module: Musicality as a Stepping Stone Towards Language Acquisition

Hirokazu Doi and Nobuo Masataka

12.1 Presumed Dissociation of Music and Language

Music is among the most ancient forms of art in human history (Masataka, 2010). The universality of music suggests that music has served some functional roles in increasing the survival odds of mankind. However, specific advantages of music conferred to human survival remain elusive. Existing theoretical and empirical studies show that music enhances group cohesion and helps emotional regulation (Tarr et al., 2014). In contrast, some theorists see no survival benefit in music, a view which is championed by Steven Pinker's famous characterization of music as "auditory cheesecake" (Pinker, 1997).

There has been a long-standing debate over whether music and language processing share the same processing stages and recruit the same neural regions. Recent studies challenge a simple dichotomy between music and language processing by showing activation of overlapping neural regions by music and language (Koelsch et al., 2002; Yu et al., 2017; but see, Peretz, et al., 2015) and impaired processing of linguistic prosody in amusic patients (Thompson et al., 2012). Thus, the linguistic ability is not as impermeable as presumed and shares part of the processing stages with music.

Interestingly, it has been reported that long-term musical training enhances language processing at various processing stages (Zhao & Kuhl, 2016), including fundamental frequency (f_0) extraction at the brainstem (Wong et al., 2007). Therefore, it seems only natural to postulate that machinery recruited in music processing has

H. Doi (✉)
School of Science and Engineering, Kokushikan University, Tokyo, Japan
e-mail: hdoi@kokushikan.ac.jp

N. Masataka (✉)
Primate Research Institute, Kyoto University, Kyoto, Japan
e-mail: masataka@pri.kyoto-u.ac.jp

S. L. Hart, D. F. Bjorklund (eds.), *Evolutionary Perspectives on Infancy*,
Evolutionary Psychology, https://doi.org/10.1007/978-3-030-76000-7_12

facilitated the evolution of language and helps in language acquisition in human infants (Masataka, 2008).

In this chapter, we first discuss the commonalities in processing stages recruited between music and language. Among the perceptual functions presumably recruited in both musical and linguistic domains, a substantial amount of knowledge has accumulated about the perception of frequency structure and rhythmic pattern. Thus, in the second section of this chapter, we provide an overview of empirical data on the developmental course of these perceptual functions in musical and linguistic domains in human infants. In the last section of this chapter, we propose a hypothesis regarding the evolutionary roots of musicality that is utilized in language acquisition in human infants today.

12.2 Commonality Between Music and Language

Language can transmit messages that are far more semantically complex than music. However, the sound stream of language shares many acoustic characteristics with musical sounds. First, vowel sounds have similar spectral structures as musical chords. When pronouncing vowels, the power of different frequencies is modified by resonance in the vocal tract, and consequently, some of the frequencies become prominent in vocal sound. Such peaks in spectral power in a few frequency ranges, called formants, define the perceived vowel category. Thus, analysis of the relative relationship between formant frequencies is required in vowel categorization. The similar ability of relative pitch perception is indispensable in music perception because the perceptual quality of musical chords is also determined by the relationship between the pitches of musical notes played simultaneously.

Second, both music and language are characterized by a hierarchical structure (Jackendoff & Lerdahl, 2006). Language is governed by grammatical rules that define the ordering of words in a hierarchical manner. Similarly, identical sound sequences are played repeatedly within a music score and are grouped into larger motifs hierarchically. Thus, the ability of perceptual grouping together with the statistical analysis of repeated sequences are valuable in analysing the global structure underlying the sound stream in both music and language.

Third, pitch contour, the temporal course of pitch change, is essential in processing both musical and linguistic sounds. Pitch contour creates a musical melody that induces vivid emotional reactions and even conveys semantic information (Koelsch et al., 2004). Likewise, linguistic prosody plays an important role in conveying emotional and semantic information (Doi et al., 2013). The human linguistic system interprets the meaning of identical sentences differently depending on the pitch contour, as in the case of pitch change at the end of utterances in declarative sentences and yes-no questions. Further, pitch contour sometimes specifies a structurally important location, e.g., phrasal boundary, in the linguistic stream.

Fourth, both music and language have a culturally transmitted rhythmic structure. Folk music in several regions has a unique metric structure that is different

from those seen in other regions of the world. Likewise, some researchers claim that language can be grouped into families according to its rhythmic structure (Nazzi & Ramus, 2003). Culture-specific patterns of auditory rhythms exert such strong influences on the perceptual system through postnatal exposure that adults have difficulty detecting slight changes and reproducing rhythmic patterns with culturally unfamiliar metric structures (Hannon & Trehub, 2005a; Collier & Wright, 1995).

12.3 Contributions of Musicality to Human Language Acquisition

Given the close similarity between music and language as described above, neural mechanisms recruited in musical processing could help in processing linguistic materials as well. Infants at the prelinguistic stage are faced with the task of analysing sound streams and grasping their underlying structure without any prior knowledge (but see Chomsky, 1965). The primary tenet of the *prosodic bootstrapping hypothesis* of language acquisition is that prosodic information, such as lexical rhythm and melodic contour in the linguistic sound stream, scaffold infants' analysis of the mother tongue. Prosodic cues are loosely related to grammatical structure. Therefore, prosodic information, a collection of musical properties of language, can help infants learn their mother tongue.

Studies on auditory perception in the fetus raise the appealing possibility that prosodic bootstrapping of language acquisition starts during the prenatal period. Fetuses are exposed to environmental sounds in the womb. However, the abdominal wall filters out high-frequency components of external sounds, and consequently, phonemic information is almost lost. Salient information contained in the filtered sound heard in the womb is mainly a prosodic pattern of language. In DeCasper et al. (1994), mothers read target rhymes each day for 4 weeks during pregnancy. Researchers then measured the fetal heart rate when hearing the target rhyme and an unfamiliar rhyme. There was a more notable deceleration in heart rate when hearing the familiar rhyme, indicating that fetuses can learn the prosodic pattern of mother's speech.

In this section, we discussed how musical functions could assist language acquisition by reviewing empirical evidence on the developmental course of these abilities. We focused especially on the perception of rhythm and frequency structure and examined whether there are any parallels between the developmental courses of these abilities in the musical and linguistic domains.

12.3.1 Perception of Frequency Structure

12.3.1.1 Musical Chord and Formant Perception

In musical chord perception, the auditory system must analyse the relationship among f_0 of simultaneously played musical notes. The difference of one-semitone in one note comprising a chord changes the perceived quality of a musical chord, from major to minor chords or vice versa. Further, adherence to certain mathematical rules creates a consonant chord, while violation of it results in a dissonant one.

There has been a debate over the innateness of chord processing ability, especially whether the preference for consonance over dissonance is acquired through cultural assimilation. Empirical studies generally favour the innateness of consonant chord preferences (Masataka, 2006). A functional magnetic resonance imaging (fMRI) study revealed differential activation of the newborn's brain in response to consonant and dissonant chords (Perani et al., 2010). An event-related potential (ERP) study indicated that newborns could discriminate between a consonant and dissonant as well as between major and minor chords (Virtala et al., 2013). These findings indicate that the infant brain is innately endowed with the capacity to analyse the pitch relationship of multiple sounds being played simultaneously.

The formant structure in vowel sounds has some resemblance to the spectral structure in musical chords. In vowel categorization, the auditory system must analyse the peak frequencies of at least three formants simultaneously. Since everyone has different exposure to f_0 depending on their gender, age, and height, the formant frequencies of identical vowel sounds differ among individuals. Despite this, identical vowels uttered by different individuals are perceived as such, a phenomenon called *speaker invariance*. Thus, what counts in vowel perception is the analysis of the relationship among formant frequencies rather than the absolute values of each formant frequency.

Behavioural studies on infant vowel categorization have shown that even neonates can discriminate between different vowel sounds (Cheour-Luhtanen et al., 1995). Vowel categorization is tuned by postnatal exposure to the mother tongue. Interestingly, some studies even argue that the tuning process starts in the womb (Moon et al., 2013). Therefore, as in the case of musical chord perception, the nascent ability of formant perception functions from the very initial stage of development.

12.3.1.2 Melody and Pitch Contour Perception

Melodic contour is created by the temporal sequence of musical notes f_0. Melody will sound the same when the notes are transposed or played in different keys (Mottron et al., 2009). This illustrates that the primary determinant of musical melody is not the absolute value of pitch but the contour of the temporal sequence of relative pitches that unfolds as the music is played.

Studies on the early development of melodic perception have shown that infants as young as 5–10 months old can detect violations of melodic contour irrespective of transposition (Trehub & Hannon, 2006). Plantinga and Trainor (2009) showed that even 2-month-olds could discriminate melodies of different songs, although the musical materials used in this study were not well controlled for low-level auditory features.

It is widely acknowledged that infants use pitch contour as a cue to analyse the grammatical structure. One example is the use of lexical stress in word segmentation (Jusczyk, 1999). Lexical stress is characterized by high pitch, long duration, large amplitude, and vowel quality. Among these acoustic cues, the human auditory system is quite sensitive to pitch change; humans can discriminate lexical stress patterns (trochaic or iambic) of nonsense syllables based on pitch-cue alone (Hoeschele & Fitch, 2016). Infants as young as 8 months can use lexical stress cues for segmenting sound streams into words (Jusczyk, 1999). In English-speaking countries, infants usually segment sound streams into units with strong-weak stress patterns. However, the 7-month-olds in Thiessen and Saffran's (2007) study switched this stress-based strategy and started extracting words with weak-strong (iambic) stress patterns after repeated exposure to word sequences of the iambic stress pattern. Thus, infants can flexibly modify their stress-based word-segmentation strategy through experience.

Another well-studied example is infants' use of pitch contour information as a cue for boundary detection. In speech, boundaries between phrases and clauses are often marked by pitch change and a long duration of syllables followed by a brief pause, which is observed cross-linguistically. Infants as young as 6 months have been shown to rely heavily on prosodic pattern for boundary detection (Seidl, 2007). Wellmann et al. (2012) investigated the acoustic characteristics that 8-month-old infants rely on in boundary detection. Pitch change or long duration alone were not sufficient for infants to detect a phrase boundary. However, the combination of long duration and pitch change enabled infants to find boundaries without pause cues. Similar results have been obtained in 6-month-olds as well (Seidl, 2007).

These studies indicate the primary importance of pitch contour in linguistic sound in order for human infants to be able to segment grammatical units. Interestingly, these studies revealed the emergence of the ability to use pitch contour in sound segmentation around 6–10 months (Thiessen & Saffran, 2007; Seidl, 2007; Wellmann et al., 2012), which roughly corresponds to the age when infants acquire the ability to process melodic contour (Trehub & Hannon, 2006). Comparisons of the developmental courses of pitch change in musical and linguistic materials also support a domain-general pattern of development (Chen et al., 2017). Indeed, such coincidence alone should not be deemed as definitive evidence, but it is quite conceivable that the maturation of identical neural mechanisms underlies the development of the ability to process pitch contour in both musical and linguistic materials.

Infants' sensitivity to pitch contour is effectively used in parental vocalizations directed towards infants. When talking to infants, adults modify their manner of speech so it differs from speech that is used with adults (Kuhl, 2007; Masataka, 2003). Such infant-directed speech is mainly characterized by high-pitch and

exaggerated intonation (Doi, 2020). High-pitched voice is effective in grabbing an infant's attention, possibly due to its emotional connotation (Corbeil et al., 2013). Likewise, exaggerated intonation makes it easier for infants to extract pitch contour, which assists infants in word segmentation and boundary detection. Thus, the domain-general ability of relative pitch perception as well as social input provided in infant-directed speech interactively scaffold infants' language acquisition (Doi, 2020; Sulpizio et al., 2018).

12.3.2 Rhythm Perception

Rhythm perception can be both objective and subjective (Iversen et al., 2009). A strong beat is often marked by large sound amplitude. At the same time, people sometimes perceive strong and weak beats in the repetition of monotonous sounds without any acoustic marks.

Rhythmicity in synchronized bodily movement is observed in as early as the neonatal stage, which is thought to reflect the activity of the central pattern generator. As for the perception of musical rhythm, an ERP study by Winkler et al. (2009) showed evidence of beat perception in neonates. In their study, neonates were exposed to a sequence of percussion sounds with a hierarchical metric structure. They were repeatedly exposed to standard sounds that lacked sound at a weak beat location. Within the sequence of standard sounds, the target sound, in which sound was omitted at a strong beat (downbeat) location, was presented with low frequency. They focused on a newborn's homologue of mismatch negativity (MMN). MMN is an ERP component usually elicited by deviant auditory sound that is presented with low frequency, embedded within a sequence of standard sounds. MMN is elicited even when a subject is not directing attention to the sounds. Thus, MMN is accepted as a reliable indicator of the ability to discriminate deviant and standard sound at the pre-attentive perceptual stage. The main finding of Winkler et al. (2009) was that the target sound elicited MMN in neonates, indicating that even neonates can discriminate weak and strong beats.

The innate ability of beat perception prepares infants to process rhythmic structures in incoming auditory information. This ability is shaped further by postnatal exposure to environmental sounds. Folk music of Eastern European countries such as Bulgaria and Macedonia has a complex metrical structure that differs from the simple metric structure in Western music. Twelve-month-olds reared in the United States, who probably had almost no experience of hearing music with a complex metre, could not detect slight changes in complex metres of Eastern European folk music (Hannon & Trehub, 2005b), while 6-month-old infants could (Hannon & Trehub, 2005a). In phoneme perception, the neural system is plastically tuned to process phoneme categories in mother tongues through postnatal experience, while losing the ability to discriminate phoneme categories that do not exist in the mother tongue (Kuhl et al., 2011). Likewise, infants lose their ability to discriminate unfamiliar types of faces, e.g. faces of different species and unfamiliar races (Pascalis

et al., 2002). The studies by Hannon and Trehub (Hannon & Trehub, 2005a, b) indicate that a similar process of perceptual tuning is at work in the development of musical rhythm; infants lose their ability to process unfamiliar rhythmic structure through postnatal exposure to their musical culture.

Linguists have raised the possibility that languages can be classified into several families according to their rhythmic structure, e.g. stress- and syllable-timed languages (Nazzi & Ramus, 2003). In stress-timed languages, such as English, the timing of the successive stressed locations in utterances are kept fairly constant, while in syllable-timed languages, such as French and Italian, syllables are uttered with constant timing. Nazzi et al. (1998) tested whether neonates could discriminate unfamiliar low-pass filtered foreign languages. Low-pass filtering eliminates high-frequency components of linguistics sound. Consequently, phonemic information is almost lost in low-pass filtered language, which makes it impossible for infants to use phoneme distribution as a clue to discriminate two languages. Interestingly, neonates in this study could discriminate unfamiliar foreign languages in different rhythmic families but not in the same rhythmic family. Thus, infants can detect rhythmic structures in languages as well as in music from the neonatal stage.

Nazzi and Ramus (2003) proposed the *rhythm bootstrapping hypothesis,* which proposes that human infants rely on rhythmic structure in language as a clue to segment grammatical units. Adults who speak syllable-timed languages use segmentation strategies in online language processing which differ from those of individuals whose mother tongue is a stress-timed language (Cutler et al., 1986). Therefore, speakers of both syllable- and stress-timed languages adopt a word segmentation strategy suited to their mother tongue. Lexical stress often signals the onset of a single word in English, but stress is less likely to be a marker of a word boundary in syllable-timed languages. Such language-specificity in word-segmentation strategies presents infant with a problem of deciding which acoustic cue to rely on in word segmentation. The rhythmic bootstrapping hypothesis suggests that the rhythmic structure of language gives infants a clue to discover the most efficient strategy, syllable- or stress-based, in word segmentation.

Most evidence for rhythm bootstrapping comes from studies on infants in an English-speaking environment. The aforementioned studies on infants learning English show that infants as young as 8 months segment words by lexical stress (Jusczyk, 1999; Thiessen & Saffran, 2007). Further, when the cues of transition probability and lexical stress are incongruent, 9-month-old infants treat lexical stress as a primary cue over transition probability in word segmentation (Thiessen & Saffran, 2007). As for syllable-timed languages, Nazzi et al. (2006) reported that French infants segmented words using a syllable-based strategy. These results, together with the early emergence of the domain-general ability to perceive rhythmic structure (Nazzi et al., 1998; Winkler et al., 2009), provide partial support to the rhythm bootstrapping hypothesis. The reason remains elusive why human infants, who are innately endowed with the ability of rhythm processing (Winkler et al., 2009), do not show evidence of language-specific segmentation strategies until around 8 months. One possible reason is that infants do not learn to associate perceived rhythmic structure with word segmentation strategy until around this age.

12.4 Evolutionary Roots of Musicality and Its Relationship with Language

Perceptual systems recruited in music processing can be used to analyse the linguistic stream and grasp its grammatical structure. The parallel development of corresponding functions in the musical and linguistic domains indicates the possibility that the maturation of domain-general functions recruited in both music and language processing assist language acquisition in human infants. This line of reasoning further raises the possibility that the evolution of musical functions has prepared the basis of the evolution of language in humans.

The burgeoning ability, or precursors, of music processing can be seen in birds, rodents, and non-human primates as well (Doi, 2020). However, their musicality does not match that of humans in its refinement. Considering this, the gap in musicality between humans and non-human species might constitute part of the reason why only humans have a sophisticated ability for language processing and speech communication. This section reviews the existing findings on musical abilities in non-human species and discusses the evolution of language from the perspective of phylogenetic roots of musicality.

12.4.1 Frequency Structure Perception

12.4.1.1 Analysis of Musical Chords and Vowel Formants

Chord perception and vowel categorization require the ability to grasp the relationships of peak frequencies. Interestingly, behavioural and electrophysiological studies have revealed close similarity in musical chord perception between non-human primates and humans (Izumi, 2000). For example, Fishman et al. (2001) measured electrophysiological responses in neurons of the primary auditory region in macaque monkeys and humans to dissonant and consonant chords. This study revealed that neurons in homologous regions in macaque monkeys' and humans' brains represent dissonance levels of musical chords.

Instrumental and vocal sounds contain harmonics of f_0. Here, harmonic means sound with the frequency of integer-multiple of f_0. When a sound composed of harmonics of the same f_0 is presented, one perceives a sound with the f_0, even when the sound lacks spectral peak in f_0. This phenomenon, called *missing fundamental*, is deemed as the expression of the superb ability of the human auditory system to analyse harmonic structure, and human infants as young as 3 months old show signs of missing fundamental perception (He & Trainor, 2009). Bendor and Wang (2005) measured activations of frequency-sensitive neurons, neurons that are activated by sound with specific frequency, in the auditory cortex in marmosets. They found a set of neurons that were activated by both pure tone with f_0 and the sound composed of harmonics of f_0 without spectral peak in f_0. These findings indicate that marmosets

possess the ability to analyse fundamental frequency in complex harmonic sound. Behavioural studies have revealed the ability to perceive missing fundamentals in other species as well (Cynx & Shapiro, 1986; Heffner & Whitfield, 1976).

These electrophysiological and behavioural studies (Bendor & Wang, 2005; Fishman et al., 2001) indicate that the basic functions for analysing the relationships of peak frequencies are phylogenetically old. Considering the spectral similarity between vowel formants and musical chords, it is possible that non-human species also possess the basic ability of vowel categorization. Direct evidence for vowel categorization has been obtained in several species (Hienz et al., 1981, 1996; Ohms et al., 2010). Among these species, songbirds show the most prominent resemblance to humans in their ability to categorize phonemes. Ohms et al. (2010) found that zebra finches could learn to discriminate vowels and generalize this discrimination to vowel sounds uttered by opposite-sex speakers despite the difference in the absolute height of formant frequencies. Thus, the findings of Ohms et al. (2010) support the view that zebra finches can analyse the relative relationship of formant frequencies in a manner closely similar to that of humans.

12.4.1.2 Analysis of Pitch Contour

In ecological settings, animals transmit many messages by modifying their vocalizations either voluntarily or involuntarily. A well-known example is the innate association between arousal level and high-pitched voice. In a highly aroused state, vocal folds vibrate at a higher frequency, which generates voices with higher f_0 in many mammalian species including humans (Bachorowski, 1999; Filippi et al., 2019; Kamiloğlu et al., 2020). In addition to the absolute height of f_0, a substantial number of studies have revealed that context-dependent messages and emotional states are encoded in the pitch contour in animal vocalizations (Briefer, 2012; Filippi et al., 2019). Therefore, the ability to analyse pitch contour must have been essential for survival.

The neural system must extract the temporal course of pitch change irrespective of the absolute pitch in the perception of pitch contour. Independence of absolute pitch and pitch contour perception is well illustrated in a phenomenon called *octave-generalization* which states that melodies sound the same when transposed by an octave. Absolute pitches of all musical notes change after transposition. Despite this, the human auditory system perceives an identical melody after transposition, which indicates strong reliance on the pitch contour, or relative pitch change, in melody processing (Mottron et al., 2009). A study by Wright et al. (2000) tested octave generalizations of musical chords in rhesus monkeys. Their main finding was that the monkeys showed signs of octave-generalization; the monkeys perceived the identical tune played in different octaves to be the same when tonal music was used as the musical material. Thus, similar to humans, rhesus monkeys also rely heavily on pitch contour rather than absolute pitch in perceiving sound sequence.

In songbirds, Spierings and ten Cate (2014) revealed that zebra finches treat prosodic patterns more heavily than structural cues in discriminating multisyllabic

sequences. Among three stress cues manipulated, i.e., pitch contour, amplitude, and duration, pitch was the most salient one to the songbirds, indicating the basic ability to discriminate rising and falling pitch contour in this species. The same group also showed that zebra finches could learn to discriminate trochaic and iambic stress patterns by pitch cue alone, but not by duration and sound amplitude cues (Spierings et al., 2017), showing the prominence of pitch contour in stress detection in zebra finches.

The number of laboratory studies reporting pitch contour perception ability in non-human species is relatively small, compared to studies on musical chord and formant perception (Fishman et al., 2001; Hienz et al., 1981, 1996; Izumi, 2000; Ohms et al., 2010). This could be because the analysis of pitch contour might be actually more difficult for non-human species than chord and formant perception; the analysis of the temporal course of pitch change requires additional functions, such as working memory, compared to the analysis of pitch relationships among simultaneously played sounds. At the same time, the prevalent use of pitch contour in the wild (Briefer, 2012; Filippi et al., 2019) indicates that lack of ecological validity in laboratory settings might have prevented researchers from finding signs of pitch contour perception in the laboratory (see Hoeschele et al., 2014, for similar discussion).

12.4.2 Rhythm Perception

Rhythm perception is closely linked to motoric functions. Indeed, human fMRI studies revealed activation of motoric regions, such as the premotor area and basal ganglia, in rhythm and beat perception (Grahn & Brett, 2007). When listening to ambient music, humans spontaneously make bodily movements in tune with the perceived beat of the music. Spontaneous entrainment to rhythm is ubiquitously observed in humans but is relatively rare among non-human species (Patel et al., 2009; Schachner et al., 2009). Even in those species showing signs of rhythmic entrainment, the temporal precision of their movement is far lower than that of humans (Hattori & Tomonaga, 2020; Patel et al., 2009).

Lack of spontaneous entrainment to rhythm does not necessarily mean the lack of ability to produce and perceive rhythmic patterns. The production of rhythmic patterns is often tested by the synchronization-continuation task (SCT). In SCT, an animal is first required to make bodily movements, usually tapping movement, in synchrony with an auditory or visual stimulus appearing at constant intervals (synchronization phase). In the continuation phase, external stimulation is eliminated and the subject must continue making bodily movements at the same pace as the synchronization phase.

An electrophysiological study using the SCT paradigm found a subgroup of neurons representing action timing in the medial premotor cortex of rhesus monkeys (Merchant et al., 2011). The firing rate of one group of neurons changed according to the elapsed time from the last action, while the other group of neurons

represented the remaining time until the next action. Thus, there are several systems for time keeping in the brain of rhesus monkeys, enabling them to make externally and internally paced actions.

However, the ability to make rhythmic movements in rhesus monkeys is not the same as that in humans. Zarco et al. (2009) compared the behaviour in SCT between rhesus monkeys and humans. After training, rhesus monkeys learned to make paced movements, but detailed analysis revealed substantial differences in the pattern of action timing between rhesus monkeys and humans. First, humans made the actions synchronously with or slightly ahead of external stimulation, indicating anticipatory preparation of timed-action. In contrast, the action timing of rhesus monkeys lagged behind external stimulation, though the action timing was faster than that in the serial reaction time task. Second, the variance of action timing was drastically larger during continuation than in the synchronization phase in rhesus monkeys at long intervals, but no such trend was found in humans. These findings raise the possibility that the mechanism for generating internally timed movements in rhesus monkeys is qualitatively different from that in humans.

Regarding the purely perceptual aspect of rhythm processing, Honing et al. (2012) measured MMN-homologue elicited by deviant stimuli in rhesus macaques using the same paradigm as Winkler et al. (2009). As explained above, a deviant sound sequence, in which sound is omitted at the downbeat location, elicits MMN in human newborns. However, the same stimuli did not reliably elicit MMN-like responses in rhesus macaques, indicating a lack of beat perception in this species.

Though several avian species show rhythmic patterns in their vocalizations, there are only mixed results on the ability of these species to perceive rhythmic structure (ten Cate & Spierings, 2019). European starlings are reported to be capable of discriminating rhythmic and arrhythmic patterns (Hulse et al., 1984). Pigeons are reported to show the ability to discriminate different metric structures only under severely limited conditions (Hagmann & Cook, 2010).

12.5 Evolutionary Roots of Language from the Perspectives of Musicality

Many species, including songbirds and non-human primates, show human-like abilities of frequency structure perception. Some species can apply this ability to process materials taken from human language (Hoeschele & Fitch, 2016; Ohms et al., 2010; Spierings & ten Cate, 2014). Further, cotton-top tamarins and rats are reported to be capable of discriminating unfamiliar languages based on prosodic cues (Ramus et al., 2000; Toro et al., 2003). Considering these, it seems likely that phylogenetically old functions are recruited in both music processing and prosodic bootstrapping of language acquisition in human infants in a domain-general manner.

In contrast to the cross-species prevalence of frequency structure perception (Briefer, 2012; Filippi et al., 2019; Fishman et al., 2001; Wright et al., 2000), few

species show the ability to perceive and produce rhythmic patterns. To obtain a full picture of the evolutionary roots of prosodic bootstrapping in language acquisition, it should be clarified how only humans acquired sophisticated rhythmic abilities. This is still a contentious field of debate, but we summarize below a tentative scenario based on existing evidence.

Schachner et al. (2009) analysed the characteristics of species that show signs of rhythmic entrainment and claimed that most of the species capable of rhythmic entrainment are also vocal learners/imitators, species that possess the ability to mimic and reproduce environmental sounds and conspecific vocalizations (Egnor & Hauser, 2004). Because both rhythmic entrainment and vocal learning require a linkage between auditory and motor systems, Patel proposed that increased functional and anatomical linkages between motor and auditory regions underlie the evolution of both vocal learning and rhythmic entrainment (Patel, 2014; Patel & Iversen, 2014). Songbirds, primates, dolphins, elephants, and bats use vocal imitation for authenticating group membership, territorial defence, adjustment of social relationships, and courtship behaviour (Coleman et al., 2007; Doi, 2020). Thus, refinement of vocal learning/imitation, hence the development of auditory-motor coupling, had conferred clear adaptive benefits to these species.

Ethnographic research on existing hunter-gatherer societies indicates the primary roles of vocal imitation in cooperative behaviour (Boyd, 2018; Lewis, 2014). Accumulation and sharing of knowledge about the surrounding environment among group members increase the odds of survival. Initiated males of the Bakaya Pygmy group achieve this by narrating their experience and knowledge through multimodal channels, including sophisticated imitation of environmental sounds as well as gestures and facial expressions (Lewis, 2014). Therefore, representing external entities by mimetic sound might have served as an efficient tool of communication in *Homos* throughout evolutionary history (Boyd, 2018). The adaptive benefit of sound-mimicking ability must have led to a closer and stronger coupling between motor and auditory regions in the human brain.

Though somewhat speculative, auditory-motor coupling may have paved the way for the emergence of fine-motor control of the vocal apparatus required in speech communication (Kearney & Guenther, 2019). The evolution of language made auditory-motor coupling even more valuable for humans, further strengthening the anatomical and functional association between these neural regions. In other words, auditory-motor coupling underlying the refinement of vocal learning/imitation prepared the basis for the evolution of speech communication. Thereafter, the survival benefit of language and speech communication in turn strengthened this coupling further.

The evolution of tight auditory-motor coupling (Rauschecker & Scott, 2009) has been driven first by the survival benefit of vocal learning/imitation and then speech communication. As a by-product of this process, humans have acquired superb ability for rhythm processing. Interestingly, neuroimaging studies indicate that motoric regions as well as auditory cortices contribute to the perception of rhythmic (Grahn & Brett, 2007) and prosodic information (Brown & Martinez, 2007; Reiterer et al., 2007; Belyk et al., 2016). Further, several studies revealed an association between

the strength of auditory-motor coupling and linguistic processing ability (Yu et al., 2017). Thus, it seems to be the case that the human brain found a way to utilize strong auditory-motor coupling to process acoustic information and hence bootstrap language acquisition during infancy.

12.6 Conclusion

Human infants are faced with the daunting task of analysing the underlying structure of linguistic sound streams. Due to the lack of any language-specific knowledge, linguistic sounds must be almost indistinguishable from music for infants. Therefore, it is natural to think that infants first apply their ability for musical processing to linguistic materials.

Indeed, existing studies point to the possibility that domain-general abilities of frequency structure and rhythm perception contribute to language acquisition during early infancy. Infants start applying these abilities to analyse linguistic sound streams almost coincidentally with the emergence of corresponding abilities in the musical domain. Such cross-domain similarity in the developmental course provides partial support for the view that domain-general functions assist language acquisition. However, this view should be validated empirically by future studies that investigate whether identical neural structures are recruited in processing musical and linguistic materials in prelinguistic infants.

The abilities for musical chord and pitch contour perception can be seen in many non-human species. Further, the species can apply these perceptual functions to analyse linguistic materials as well. In contrast, animals do not match humans in their ability to perceive and produce rhythmic patterns. Considering these, what separates humans from non-human species in terms of language evolution seems to be the emergence of tight coupling between auditory and motor regions in humans that engendered both speech communication and refined rhythmic ability.

The emergence of language made auditory-motor coupling more beneficial for humans. The strong auditory-motor coupling has brought up musical and linguistic abilities to even higher levels in humans, thereby enabling infants to use these domain-general functions to analyse linguistic sound streams in the process of language acquisition.

References

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*(2), 53–57.

Belyk, M., Pfordresher, P. Q., Liotti, M., & Brown, S. (2016). The neural basis of vocal pitch imitation in humans. *Journal of Cognitive Neuroscience, 28*(4), 621–635.

Bendor, D., & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature, 436*, 1161–1165.

Boyd, B. (2018). The evolution of stories: From mimesis to language, from fact to fiction. *WIREs Cognitive Science, 9*(1), 1–16.

Briefer, E. F. (2012). Vocal expression of emotions in mammals: Mechanisms of production and evidence. *Journal of Zoology, 288*(1), 1–20.

Brown, S., & Martinez, M. J. (2007). Activation of premotor vocal areas during musical discrimination. *Brain and Cognition, 63*(1), 59–69.

Chen, A., Stevens, C. J., & Kager, R. (2017). Pitch perception in the first year of life, a comparison of lexical tones and musical pitch. *Frontiers in Psychology*, *8*, Article 297.

Cheour-Luhtanen, M., Alho, K., Kujala, T., Sainio, K., Reinikainen, K., Renlund, M., Aaltonen, O., Eerola, O., & Näätänen, R. (1995). Mismatch negativity indicates vowel discrimination in newborns. *Hearing Research, 82*(1), 53–58.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Coleman, S. W., Patricelli, G. L., Coyle, B., Siani, J., & Borgia, G. (2007). Female preferences drive the evolution of mimetic accuracy in male sexual displays. *Biology Letters, 3*(5), 463–466.

Collier, G. L., & Wright, C. E. (1995). Temporal rescaling of simple and complex ratios in rhythmic tapping. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 602–627.

Corbeil, M., Trehub, S. E., & Peretz, I. (2013). Speech vs. singing: Infants choose happier sounds. *Frontiers in Psychology*, *4*, Article 372.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25*(4), 385–400.

Cynx, J., & Shapiro, M. (1986). Perception of missing fundamental by a species of songbird (Sturnus vulgaris). *Journal of Comparative Psychology, 100*(4), 356–360.

DeCasper, A. J., Lecanuet, J.-P., Busnel, M.-C., Granier-Deferre, C., & Maugeais, R. (1994). Fetal reactions to recurrent maternal speech. *Infant Behavior & Development, 17*(2), 159–164.

Doi, H. (2020). Social scaffolding of vocal and language development. In N. Masataka (Ed.), *The origins of language revisited: Differentiation from music and the emergence of neurodiversity and autism* (pp. 115–137). Springer Nature Singapore.

Doi, H., Fujisawa, T. X., Kanai, C., Ohta, H., Yokoi, H., Iwanami, A., Kato, N., & Shinohara, K. (2013). Recognition of facial expressions and prosodic cues with graded emotional intensities in adults with Asperger syndrome. *Journal of Autism and Developmental Disorders, 43*(9), 2099–2113.

Egnor, S. E. R., & Hauser, M. D. (2004). A paradox in the evolution of primate vocal learning. *Trends in Neuroscience, 27*(11), 649–654.

Filippi, P., Hoeschele, M., Spierings, M., & Bowling, D. L. (2019). Temporal modulation in speech, music, and animal vocal communication: Evidence of conserved function. *Annals of New York Academy of Science, 1453*(1), 99–113.

Fishman, Y. I., Volkov, I. O., Noh, M. D., Garell, P. C., Bakken, H., Arezzo, J. C., Howard, M. A., & Steinschneider, M. (2001). Consonance and dissonance of musical chords: Neural correlates in auditory cortex of monkeys and humans. *Journal of Neurophysiology, 86*(6), 2761–2788.

Grahn, J. A., & Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *Journal of Cognitive Neuroscience, 19*(5), 893–906.

Hagmann, C. E., & Cook, R. G. (2010). Testing meter, rhythm, and tempo discriminations in pigeons. *Behavioural Processes, 85*(2), 99–110.

Hannon, E. E., & Trehub, S. E. (2005a). Metrical categories in infancy and adulthood. *Psychological Science, 16*(1), 48–55.

Hannon, E. E., & Trehub, S. E. (2005b). Tuning in to musical rhythms: Infants learn more readily than adults. *Proceedings of the National Academy of Sciences of the United States of America, 102*(35), 12639–12643.

Hattori, Y., & Tomonaga, M. (2020). Rhythmic swaying induced by sound in chimpanzees (*Pan troglodytes*). *Proceedings of the National Academy of Sciences of the United States of America, 117*(2), 936–942.

Heffner, H., & Whitfield, I. C. (1976). Perception of the missing fundamental by cats. *The Journal of the Acoustical Society of America, 59*, 915.

He, C., & Trainor, L. J. (2009). Finding the pitch of the missing fundamental in infants. *Journal of Neuroscience, 29*, 7718–8822.

Hienz, R. D., Aleszczyk, C. M., & May, B. J. (1996). Vowel discrimination in cats: Acquisition, effects of stimulus level, and performance in noise. *Journal of Acoustical Society of America, 99*, 3656–3668.

Hienz, R. D., Sachs, M. B., & Sinnott, J. M. (1981). Discrimination of steady-state vowels by blackbirds and pigeons. *Journal of Acoustical Society of America, 70*, 699–706.

Hoeschele, M., & Fitch, W. T. (2016). Phonological perception by birds: Budgerigars can perceive lexical stress. *Animal Cognition, 19*(3), 643–654.

Hoeschele, M., Merchant, H., Kikuchi, Y., Hattori, Y., & ten Cate, C. (2014). Searching for the origins of musicality across species. *Philosophical Transactions of the Royal Society. B, Biological Sciences, 370*(1664), 20140094.

Honing, H., Merchant, H. H., Gábor, P., Prado, L., & Bartolo, R. (2012). Rhesus monkeys (*Macaca mulatta*) detect rhythmic groups in music, but not the beat. *PLoS One, 7*(12), e51369.

Hulse, S. H., Humpal, J., & Cynx, J. (1984). Discrimination and generalization of rhythmic and arrhythmic sound patterns by European starlings (*Sturnus vulgaris*). *Music Perception, 1*(4), 442–464.

Iversen, J. R., Repp, B. H., & Patel, A. D. (2009). Top-down control of rhythm perception modulates early auditory responses. *Annals of the New York Academy of Sciences, 1169*, 58–73.

Izumi, A. (2000). Japanese monkeys perceive sensory consonance of chords. *Journal of the Acoustical Society of America, 108*(6), 3073–3078.

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

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences, 3*(9), 323–328.

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review, 27*(2), 237–265.

Kearney, E., & Guenther, F. H. (2019). Articulating: The neural mechanisms of speech production. *Language, Cognition and Neuroscience, 34*, 1214–1229.

Koelsch, S., Gunter, T. C., Cramon, D. Y., Zysset, S., Lohmann, G., & Friederici, A. D. (2002). Bach speaks: A cortical "language-network" serves the processing of music. *NeuroImage, 17*(2), 956–966.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: Brain signatures of semantic processing. *Nature Neuroscience, 7*, 302–307.

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science, 10*(1), 110–120.

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2011). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America, 100*(15), 9096–9101.

Lewis, J. D. (2014). *BaYaka Pygmy multi-modal and mimetic communication traditions*. Oxford University Press, 19(7), 77–91.

Masataka, N. (2003). How infant-directed speech influences infant vocal development. In N. Masataka (Ed.), *The onset of language* (pp. 126–156). Cambridge University Press.

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

Masataka, N. (2008). Implication of the human musical faculty for evolution of language. In N. Massataka (Ed.), *The origins of language: Unraveling evolutionary forces* (pp. 133–152). Springer Nature.

Masataka, N. (2010). Were musicians as well as artists in the ice age caves likely with autism Spectrum disorder? A neurodiversity hypothesis. In N. Masataka (Ed.), *The origins of*

language revisited: Differentiation from music and the emergence of neurodiversity and autism (pp. 323–345). Springer Nature Singapore.

Merchant, H., Zarco, W., Pérez, O., Prado, L., & Bartolo, R. (2011). Measuring time with different neural chronometers during a synchronization-continuation task. *Proceedings of the National Academy of Sciences of the United States of America, 108*(49), 19784–19789.

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experienced in utero affects vowel perception after birth: A two-country study. *Acta Paediatrica, 102*(2), 156–160.

Mottron, L., Dawson, M., & Soulières, I. (2009). Enhanced perception in savant syndrome: Patterns, structure and creativity. *Philosophical Transactions of Royal Society of London, Series B: Biological Science, 364*(1522), 1385–1391.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 756–766.

Nazzi, T., Iakimova, G., Bertoncini, J., Frédonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language, 54*(3), 283–299.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication, 41*(1), 233–243.

Ohms, V. R., Gill, A., van Heijningen, C. A. A., Beckers, G. J. L., & ten Cate, C. (2010). Zebra finches exhibit speaker-independent phonetic perception of human speech. *Proceedings of the Royal Society of London Series B: Biological sciences, 277*(1684), 1003–1009.

Pascalis, O., de Haan, M., & Nelson, C. A. (2002). Is face processing species-specific during the first year of life? *Science, 296*(5571), 1321–1323.

Peretz, I., Vuvan, D., Lagrois, M. E., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 370*(1664), 20140090.

Patel, A. D. (2014). The evolutionary biology of musical rhythm: Was Darwin wrong? *PLoS Biology, 12*(3), e1001821.

Patel, A. D., & Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: The Action Simulation for Auditory Prediction (ASAP) hypothesis. *Frontiers in Systems Neuroscience*, *8*, Article 57.

Patel, A. D., Iversen, J. R., Bregman, M. R., & Schulz, I. (2009). Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current Biology, 19*(10), 827–830.

Perani, D., Saccuman, M. C., Scifo, P., Spada, D., Andreolli, G., Rovelli, R., Baldoli, C., & Koelsch, S. (2010). Functional specializations for music processing in the human newborn brain. *Proceedings of the National Academy of Sciences of the United States of America, 107*(10), 4758–4763.

Pinker, S. (1997). *How the mind works*. WW. Norton & Company.

Plantinga, J., & Trainor, L. J. (2009). Melody recognition by two-month-old infants. *The Journal of the Acoustical Society of America, 125*, EL58.

Reiterer, S., Erb, M., Grodd, W., & Wildgruber, D. (2007). Cerebral processing of timbre and loudness: fMRI evidence for a contribution of Broca's area to basic auditory discrimination. *Brain Imaging and Behavior, 2*(1), 1–10.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science, 288*(5464), 349–351.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience, 12*(6), 718–724.

Schachner, A., Brady, T. F., Pepperberg, I. M., & Hauser, M. D. (2009). Spontaneous motor entrainment to music in multiple vocal mimicking species. *Current Biology, 19*(10), 831–836.

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language, 57*(1), 24–48.

Spierings, M. J., & ten Cate, C. (2014). Zebra finches are sensitive to prosodic features of human speech. *Proceedings of the Royal Society. Biological Sciences, 281*(1787), 20140480.

Spierings, M., Hubert, J., & ten Cate, C. (2017). Selective auditory grouping by zebra finches: Testing the iambic–trochaic law. *Animal Cognition, 20*(4), 665–675.

Sulpizio, S., Doi, H., Bornstein, M. H., Cui, J., Esposito, G., & Shinohara, K. (2018). fNIRS reveals enhanced brain activation to female (versus male) infant directed speech (relative to adult directed speech) in young human infants. *Infant Behavior and Development, 52*, 89–96.

Tarr, B., Launay, J., & Dunbar, R. I. M. (2014). Music and social bonding: "Self-other" merging and neurohormonal mechanisms. *Frontiers in Psychology, 5*, Article 1096.

ten Cate, C., & Spierings, M. (2019). Rules, rhythm and grouping: Auditory pattern perception by birds. *Animal Behaviour, 151*, 249–257.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3*(1), 73–100.

Thompson, W. F., Marin, M. M., & Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 109*(46), 19027–19032.

Trehub, S. E., & Hannon, E. E. (2006). Infant music perception: Domain-general or domain-specific mechanisms? *Cognition, 100*(1), 73–99.

Toro, J. M., Trobalon, J. B., & Sebastián-Gallés, N. (2003). The use of prosodic cues in language discrimination tasks by rats. *Animal Cognition, 6*(2), 131–136.

Virtala, P., Huotilainen, M., Partanen, E., Fellman, V., & Tervaniemi, M. (2013). Newborn infants' auditory system is sensitive to Western music chord categories. *Frontiers in Psychology*, *4*, Article 492.

Wellmann, C., Holzgrefe, J., Truckenbrodt, H., Wartenburger, I., & Höhle, B. (2012). How each prosodic boundary cue matters: Evidence from German infants. *Frontiers in Psychology, 3*, 580.

Winkler, I., Háden, G. P., Ladinig, O., Sziller, I., & Honing, H. (2009). Newborn infants detect the beat in music. *Proceedings of the National Academy of Sciences of the United States of America, 106*(7), 2468–2471.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10*(4), 420–422.

Wright, A. A., Rivera, J. J., Hulse, S. H., Shyan, M., & Neiworth, J. J. (2000). Music perception and octave generalization in rhesus monkeys. *Journal of Experimental Psychology: General, 129*(3), 291–307.

Yu, M., Xu, M., Li, X., Chen, Z., Song, Y., & Liu, J. (2017). The shared neural basis of music and language. *Neuroscience, 357*, 208–219.

Zarco, W., Merchant, H., Prado, L., & Mendez, J. C. (2009). Subsecond timing in primates: Comparison of interval production between human subjects and rhesus monkeys. *Journal of Neurophysiology, 102*(6), 3191–3202.

Zhao, T. C., & Kuhl, P. K. (2016). Musical intervention enhances infants' neural processing of temporal structure in music and speech. *Proceedings of the National Academy of Sciences of the United States of America, 113*(19), 5212–5217.