Developmental Science 10:1 (2007), pp 35–39 DOI: 10.1111/j.1467-7687.2007.00561.x

Music, evolution and language

Nobuo Masataka

Japan Science Technology Agency, Saitama, Japan and Kyoto University, Inuyama, Japan

Abstract

Darwin (1871) noted that the human musical faculty 'must be ranked amongst the most mysterious with which he is endowed'. Indeed, previous research with human infants and young children has revealed that we are born with variable musical capabilities. Here, the adaptive purpose served by these differing capabilities is discussed with reference to comparative findings regarding the acoustic behavior of nonhuman primates. The findings provide evidence supporting Darwin's hypothesis of an intermediate stage of human evolutionary history characterized by a communication system that resembles music more closely than language and possibly acting as a precursor for both current language and music.

Introduction

Although the debate on the origin of human language has a long history that continues up to the present, there is consensus concerning the fact that the language system arose by means of natural selection, presumably because more accurate communication helped early humans survive and reproduce. However, the evolutionary significance of music remains open to question. Pinker (1997) stated that music itself played no adaptive role in human evolution. He suggested that it was 'auditory cheesecake', a byproduct of natural selection that just happened to 'tickle the sensitive spots' of other truly adaptive functions, such as the rhythmic bodily movement of walking and running, the natural cadences of speech, and the brain's ability to make sense of a cacophony of sounds. However, a number of researchers disagree with this argument (e.g. Christiansen & Kirby, 2003; Wallin, Merker & Brown, 2000). These authors argue that music clearly had an evolutionary role and point to its universality. From a developmental perspective, it is interesting to note that even very young infants respond strongly to music. Perhaps this indicates that music is somehow hardwired into the human brain. But if infants' musical ability is the result of Darwinian natural selection, in what way did it make humans more fit?

Numerous findings concerning the evolution of language as a system have recently been reported in the area of cognitive science, particularly developmental cognitive science, providing material for the presentation of a concrete scenario (for a review see Masataka, 2003, in press). These findings indicate that children learn the subcomponents of the language system one after another, with the time course remaining largely consistent regardless of the language system being acquired. This suggests that although children have to learn each subcomponent, the activity of learning itself is genetically preprogrammed. However, linguists tend to view the evolutionary onset of each component as an abrupt phenomenon (e.g. Pinker, 1994). In contrast, developmental cognitive scientists view the attainment of each subcomponent as deeply intertwined. In line with this, questions arise as to how each subcomponent evolved into its present form, what the previous forms were, as well as when and how several subcomponents came to be related to one another. Answering each of these questions would help reconstruct language evolution, and we believe explain the presence of infants' puzzling cognitive musical ability. The present review is a preliminary attempt to provide a conceptual perspective on the above questions with the aim of helping to unravel the implications of the human musical faculty, particularly with reference to the evolution of language.

Coos in human infants and Japanese macaque adults

Of the various early language development milestones, the earliest is perceptual competence related to speech sound discrimination in typically developing infants

Address for correspondence: Nobuo Masataka, Primate Research Institute, Kyoto University, Inuyama, Aichi, Japan; e-mail: masataka@pri.kyoto-u.ac.jp

exposed to spoken language (Werker & Vaoloumanos, 2000). At birth, the newborn has the ability to distinguish virtually all sounds used in all languages, at least when the sounds are presented in isolation. The newborn produces no speech sounds, however. Although speech-like sounds gradually emerge during the first year of life, the general consensus is that two discrete stages are recognized in the early process of spoken language production.

The first stage begins with vowel-like monosyllabic sounds (coos) at 6–8 weeks of age. The learning process that preverbal infants must undertake then is phonation. The relevant learning occurs between 3 and 4 months of life, through turn-taking with caregivers. The timing and quality of adult vocal responses affect the social vocalizations of infants around this age. Interestingly, such interaction has been reported in a nonhuman primate, the Japanese macaque (*Macaca fuscata*). Animals of this species utter similar coos to maintain vocal contact. When my colleague and I (Sugiura & Masataka, 1995) observed exchanges of the vocalizations in free-ranging populations, the temporal patterns of occurrence of the intercall intervals between two consecutive coos during vocal interaction were similar to those obtained in human mother–infant dyads; after a monkey cooed spontaneously, it remained silent for a short interval, and if no response was heard from the other monkeys, then the monkey would often coo again to address the other group members.

While one of the striking aspects of human spoken language certainly lies in the importance of auditory feedback during development, a wide variety of studies have presented evidence that such macaque vocalizations also undergo similar modification as a function of social context. This evidence can be summarized in two major categories: acoustic variation between social groups and social convergence. My colleagues and I recently reported the results of cross-sectional and longitudinal comparisons of the acoustic features of the coos between two groups, both of which derived from the same local population but had been separated for more than 34 years (Tanaka, Sugiura & Masataka, 2006). When the frequencies of the fundamental frequency element (Fo) in the vocalizations were recorded from more than 50 individuals varying in age from 6 months to 18 years, small but significant differences were consistently noted between the groups of animals older than 1 year. Such differences were not found in younger individuals, suggesting that they arise from learning. While it is still difficult to completely rule out genetic factors, we assume that such differences reflect an underlying modification of a fixed template, which is similar to but more subtle than what has been reported on the vocal flexibility of other animals such as songbirds. Supposedly, this could increase the expressive potential of a vocal communication system and might be crucial for advertising and maintaining social group membership, committing to a current alliance, or indicating the receipt of distant calls.

As for the convergence of acoustic features, Sugiura (1998) reported in free-ranging Japanese macaques that a coo sound with a rising pitch contour is likely to be responded to by another coo with a rising contour, and vice versa. This sort of vocal interaction was shown to occur between individuals affiliated although unrelated to one another, hence functioning to maintain and even strengthen the relationship between them. Subsequently, Sugiura demonstrated using a playback experiment that the animals matched the acoustic features of response 'coo' vocalizations to the eliciting stimulus 'coo' on a short timescale. Since the publication of this study, similar phenomena have been reported in New World monkeys and great apes (Masataka, 2003). In both cases, flexible variability occurred acoustically in the vocalizations with respect to the pattern of temporal organization of the frequency modulation of their tonal elements, particularly Fo. When attempting to vocalize spontaneously, individuals are required to choose an acoustic variation of a single type of sound from several such variations. When attempting to respond to the sound, an opportunity with the identical option is provided with the attempting individual. However, it should be noted that such vocal matching is observed exclusively between adult individuals, and not in interactions involving juveniles or infants. Although longitudinal data are still required, some kind of social experience seems to be necessary for the acquisition of this ability.

Prosodic communication in human infants

In humans, on the other hand, similar abilities have already been observed in preverbal infants before they are able to produce well-articulated sounds. Halliday (1975) made the first systematic attempt at examining the significance of within-individual variability of the acoustic flexibility of such vocalizations, reporting voluntarily variation as a means of signaling different communicative functions. He found the rising pitch contour of human infant coos to be produced in association with 'pragmatic' functions, such as requests for objects, and the falling pitch to be associated with 'mathetic' functions such as labeling. It is also notable that, in humans, caregivers intuitively encourage infants to perform vocal matching of this sort, whereas no such encouragement is observed in nonhuman primates. This parenting style is conventionally referred to as motherese, a speech style characteristic of adults when addressing infants and young children. Particularly, prosodic modification is considered

to be cross-culturally universal, whereby speech takes on an elevated and exaggerated pitch contour whether caregivers are aware of it or not.

As for the evolution of motherese, two functional roles have been pointed out (Werker & McLeod, 1989). The primary role is referred to as the 'attention-getting property'. Even newborns exhibit a strong tendency to direct their attention more toward speech sounds with exaggerated pitch excursions than to those without this feature. The second functional category is referred to as 'affective salience'. In a series of experimental studies, participating infants looked at the speaker from whom the motherese stimulus was delivered with more positive affective responsiveness than to the speaker from whom non-motherese sounds were delivered. Taken together, affective salience and the attention-getting property suggest a linguistic benefit for preverbal infants (Masataka, 1992). In my observation, the pitch contour of the coo sounds of 6-month-old infants matches that of spontaneous maternal utterances when the coo occurs in response to the maternal utterance. However, this phenomenon was confirmed only when the maternal utterance was provided with motherese features. When the utterance was not significantly modified, the pitch contour of the responding coos was determined regardless of the suprasegmental features of the preceding maternal speech. Prosodically, motherese works through the facilitation of language learning by preverbal infants. Falk (2004) hypothesizes that as the brain size of early hominids increased, human infants became unable to cling onto their mothers. The hominid female is assumed to have responded to this situation by developing motherese so that interaction with her infant could become more coordinated, eventually providing the infant with opportunities to acquire the capability to learn much more flexible vocal usage than would otherwise have been provided. This notion is supported by the social brain hypothesis (Dunbar, 1993), which suggests that larger human brain sizes and language both evolved as a response to increasing group sizes in our primate ancestors.

The evolution of 'singing' behavior

The onset of the second stage in the early process of spoken language production is around 8 months of age, when speech-like vocalizations in infancy culminate in the emergence of babbling. Around the same time, infants become able to temporally retain auditory information, associating stored patterns of sounds with the patterns of sounds they produce. Unlike the sounds that infants produce before this stage, babbling consists of well-formed syllables that have adult-like spectral and temporal properties. One month later, the babbling is articulated with a rapid formant transition duration in a relatively short syllable. Results of various acoustical analyses with the vocalizations demonstrated that there is a significant continuity between the sound systems of babbling and early speech, and that units present in babbling are utilized later in natural spoken languages (for a review see Masataka, 2003). Finally, at approximately one year after birth, first words are observed.

In nonhuman primates, no vocalizations are as well articulated as the babbling of 9-month-old human infants. No forms of multisyllabic utterances occur in monkeys or prosimians. However, long-distance calls produced by apes are commonly characterized by pure tonal notes, stereotyped phrases, biphasic notes, accelerando in note rhythm, and possibly a slow-down near the end of the phrase. They are acoustically similar to the multisyllabic sounds that human infants as young as 8 months of age produce in a poorly articulated manner. This has been investigated most intensively in gibbons. Darwin (1871) had already noted the importance of this similarity and argued that 'primeval man, or rather some early progenitor of man, probably first used his voice in producing true musical cadences, that is in singing, as do some of the gibbon-apes at the present day' (p. 133), because all gibbon species use a variety of different note types as a repertoire of their 'songs'. Recent results from analyses of gibbon calls have presented further evidence for the plausibility of this hypothesis about the evolution of human language (Haimoff, 1986). There is one gibbon species that is unique in that its song repertoire does not appear to include sex-specific note types. In this species, all types of notes occur in the short phrases of both males and females. There is another group of gibbons that represents the other extreme of the spectrum, showing the highest degree of sex-specificity in their note type repertoire, with males and females of this species both producing several note types, each of which is not normally produced by conspecifics of the opposite sex. In all remaining gibbon species, adult males and females share certain components of the note type repertoire, but also use some sex-specific notes. When arranging the song characteristics of various gibbon species linearly according to the sex-specificity of the song repertoire, the results reveal duets in which both pair partners sing virtually identical duet contributions to pairs in which the repertoires of both sexes overlap partially, and finally, to pairs in which the repertoires are completely sex-specific, whereby each sex confines its vocalizations to only one part of the whole song. This linear arrangement is interpreted as representing an evolutionary trend from solo singing to full partner dependence and increasing 'song-splitting' (Wickler & Seibt, 1982).

Furthermore, male solo as well as duet song bouts were found to occur in the mated pairs of one species (Geissmann, 2002). The majority of these solo songs were heard approximately 2 hours after dawn but approximately 2 hours earlier than duet songs. In another group, however, the first peak of singing activity occurred at, or even before, sunrise, and this time the males produced solo songs from their sleeping tree. The second peak occurred one to several hours later, after the first feeding bout, and then the females would usually join the males in duet songs. There are two additional species that are exceptional in that pair partners sing solo songs only, suggesting that these species represent a stage before the evolution of duetting. Under other circumstances, however, well-coordinated duets and their patterns of duetting are the most elaborate forms of singing produced by gibbon species. Thus, an alternative view is that solo singing in these two species was derived secondarily from duet singing. This evolutionary process is designated 'duet-splitting' (Geissmann, 2002), in which the contributions of pair partners were split into temporally segregated solo songs.

An intermediate role for music in the evolution of language?

While the production of multisyllabic calls by apes has conventionally been termed 'singing', most characteristics provided with the sounds are also recognized in singing by modern humans, regardless of the culture in which they live (Boulez, 1971). It is therefore possible that the loud calls of early hominids shared the above characteristics with apes, providing the basis from which current human language evolved. Given this assumption, the reason why multisyllabic sound utterances were adopted as media to embody language competence in our ancestors is no longer puzzling, and moreover, the successive evolution of duet-splitting and song-splitting in gibbon singing might provide us with a conceptual framework to reconstruct the transitive process from singing to real speaking. According to the hypothesis described in this paper, duet-splitting occurred in mated pairs who first sang a sequence of songs together followed by song-splitting. Once a certain component is vocalized independently from the other parts, the result will no longer be singing. Moreover, if this is executed under voluntary motor control and the influence of auditory feedback, the result could almost be construed as speech-like behavior. This possibility should be explored in the future.

At an early stage of development, human infants perceive speech sounds as music, and are likely to attend to the melodic and the rhythmic aspects of speech. Similar findings have been reported in some nonhuman primates (Masataka, in press). Based on such common cognitive properties, human infants are thus enabled to acquire spoken languages; their predisposition to the properties of spoken language passed on during human evolution. Consequently they exhibit a cross-culturally universal, particular pattern of discrimination when music stimuli are presented (Trainor & Heinmiller, 1998; Trainor, Tsang & Cheung, 2002; Masataka, 2006). Although it has been argued that the early pattern of language discrimination and recognition reflects an innate, language-specific predisposition unique to humans, this assumption has recently been challenged (Fishman, Volkov, Noh, Garell, Bakken, Arezzo, Howard & Steinschneider, 2001). While in humans the neural mechanism for language processing is believed to be located in Wernicke's area, there is evidence that the fixation of a gene (*FOXP2*) expressing in Broca'a area occurred during the last 200,000 years of human history (Enard, Przeworski, Fisher, Lai, Wiebe, Kitano, Monaco & Paabo, 2002). Deficits of this gene are associated with the occurrence of motor speech disorders (Watkins, Dronkers & Vergha-Khadem, 2002). This suggests that the evolution of the region for language processing occurred earlier than the system including Broca's area. In the human brain as well as in the brains of most nonhuman primates, the auditory areas consist of the primary auditory cortex and auditory association area (the supratemporal gyrus). Further, the neural network that projects from the inner ear to the primary auditory cerebral cortex is formed without any auditory input in the brain of both humans and nonhuman primates. However, post-processing neurons in humans develop with learning by proper neural input whereas in nonhuman primates this opportunity for modification is extremely limited. In humans, the learning period is thought to occur below 5 to 6 years of age. Reducing auditory signals during this critical language-learning period can limit a child's potential for developing an effective communication system.

Human infants are innately predisposed to discover the particular patterned input of phonetic and syllabic units, and only as a result of this can post-processing neurons develop. These units represent the particular patterns of the input signal, and in humans they correspond to the temporal and hierarchical grouping and rhythmical characteristics of natural spoken language phonology. Moreover, a similar cognitive mechanism is thought to be more extensively shared with some nonhuman primate species than has been assumed so far. The most plausible explanation for this sharing is perhaps similarity in the communication systems of some nonhuman primates and our ancestors, which appear to resemble music rather than language in its present form.

References

Boulez, P. (1971). *Boulez on music today*. London: Faber and Faber.

Christiansen, M.H., & Kirby, S. (2003). *Language evolution*. Oxford: Oxford University Press.

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

Dunbar, R.I.M. (1993). Coevolution of neocortex size, group size and language in humans. *Behavioral and Brain Sciences*, **16**, 681–735.

Enard, W., Przeworski, W., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., & Paabo, S. (2002). Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature*, **418**, 869–872.

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

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**, 2761–2788.

Geissmann, T. (2002). Duet-splitting and the evolution of gibbon songs. *Biological Review*, **77**, 57–76.

Haimoff, E.H. (1986). Convergence in the duetting of monogamous Old World primates. *Journal of Human Evolution*, **15**, 51–59.

Halliday, M.A.K. (1975). *Learning how to mean: Explorations in the development of language*. London: Edward Arnold.

Masataka, N. (1992). Pitch characteristics of Japanese maternal speech to infants. *Journal of Child Language*, **19**, 213–223.

Masataka, N. (2003). *The onset of language*. Cambridge: Cambridge University Press.

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

Masataka, N. (in press). *The origins of language reconsidered*. Heidelberg: Springer.

Pinker, S. (1994). *The language instinct*. New York: HarperCollins.

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

Sugiura, H. (1998). Matching of acoustic features during vocal exchange of coo calls in Japanese macaques. *Animal Behaviour*, **55**, 673–687.

Sugiura, H., & Masataka, N. (1995). Temporal and acoustic flexibility in vocal exchange of coo calls in Japanese monkeys (*Macaca fuscata*). In E. Zimmermann, J.D. Newman, & U. Jurgens (Eds.), *Current topics in primatology* (pp. 121–140). London: Plenum.

Tanaka, T., Sugiura, H., & Masataka, N. (2006). Cross-sectional and longitudinal studies of group differences in acoustic features of coo calls in two groups of Japanese macaques. *Ethology*, **112**, 7–21.

Trainor, L.J., & Heinmiller, B.M. (1998). The development of evaluative responses to music: infants prefer to listen to consonance over dissonance. *Infant Behavior and Development*, **21**, 77–88.

Trainor, L.J., Tsang, C.D., & Cheung, V.H.W. (2002). Preference for sensory consonance in 2- and 4-month-old infants. *Music Perception*, **20**, 187–194.

Wallin, N.L., Merker, B., & Brown, S. (2000). *The origins of music*. Cambridge, MA: MIT Press.

Watkins, K.E., Dronkers, N.F., & Vergha-Khadem, F. (2002). Behavioral analysis of an inherited speech and language disorder: comparison with acquired aphasia. *Brain*, **125**, 452–464.

Werker, J.F., & McLeod, P.J. (1989). Infant preference for both male and female infant-directed talk: a developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology*, **43**, 320–346.

Werker, J.F., & Vouloumanos, A. (2000). Who's got rhythm? *Science*, **288**, 280–281.

Wickler, W., & Seibt, U. (1982). Song splitting in the evolution of duetting. *Zeitschrift fur Tierpsychologie*, **59**, 127–140.