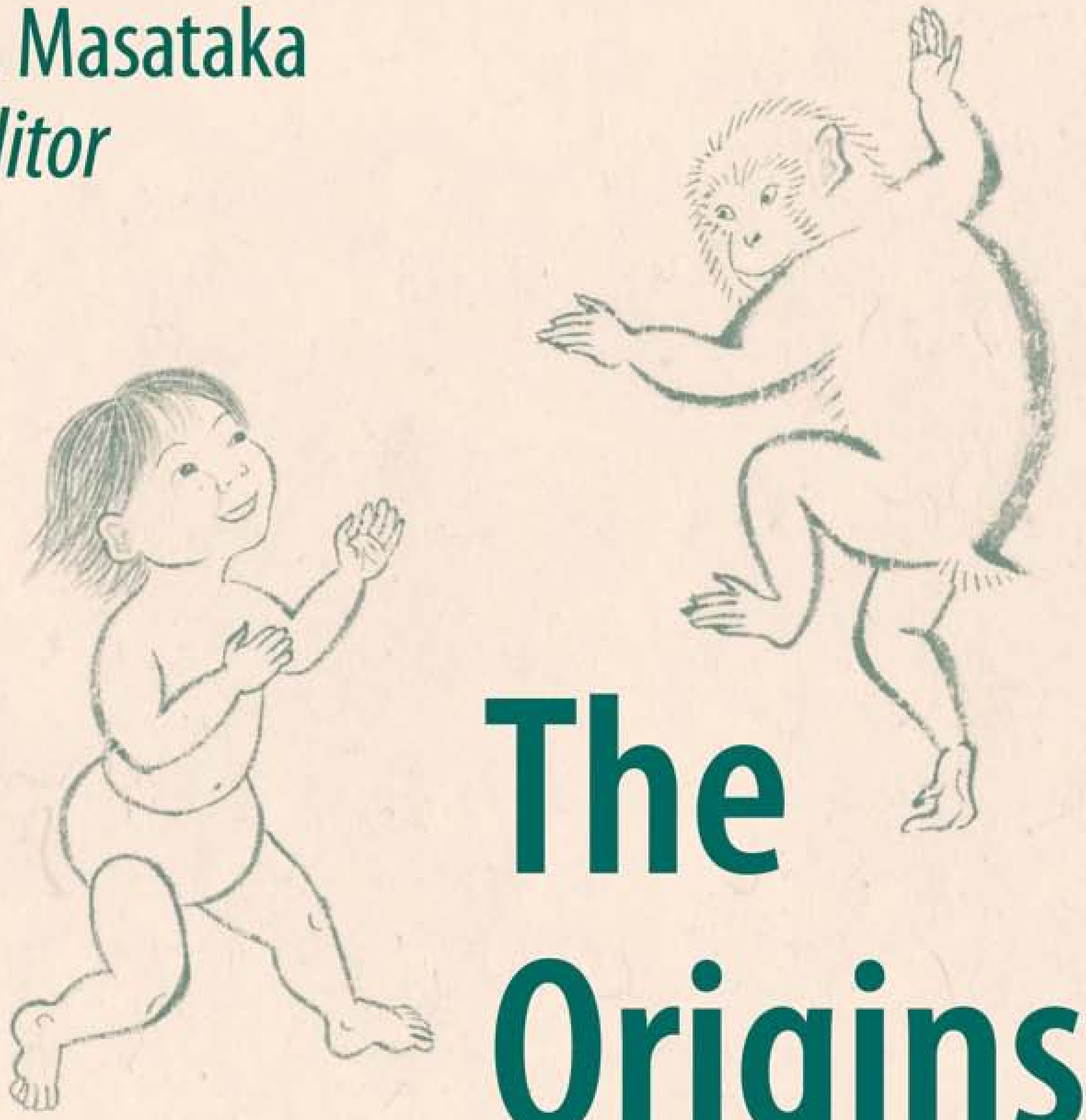
N. Masataka

Editor



The Origins of Language

Unraveling Evolutionary Forces

Springer

Nobuo Masataka (Ed.)

The Origins of Language

Unraveling Evolutionary Forces

Nobuo Masataka (Ed.)

The Origins of Language

Unraveling Evolutionary Forces

Springer

Nobuo Masataka
Professor, Primate Research Institute, Kyoto University
41 Kanrin, Inuyama, Aichi 484-8506, Japan

Cover: “Man Meets Monkey” drawn by Motoko Masataka

ISBN 978-4-431-79101-0 Springer Tokyo Berlin Heidelberg New York
e-ISBN 978-4-431-79102-7

Library of Congress Control Number: 2008928680

Printed on acid-free paper

Springer is a part of Springer Science+Business Media
springer.com

Typesetting: SNP Best-set Typesetter Ltd., Hong Kong
Printing and binding: Hicom, Japan

Preface

Debate on the origins of language has a long—and primarily speculative—history. Perhaps its most significant milestone occurred in 1866, when the Société de Linguistique de Paris banned further papers on the subject, because fossil records could provide no evidence concerning linguistic competence. This view has persisted until recently, with investigators who deal with language empirically remaining largely on the sidelines.

Contemporary developments in cognitive science, however, indicate that human and nonhuman primates share a range of behavioral and physiological characteristics (e.g., perceptual and computational) that speak to this issue of language origins. Rather than indicating a discontinuity between humans and other animals, studies concerning communicative, neurological, and social aspects of language behavior suggest that the view of language as determined by biologically innate abilities in conjunction with exposure to language in an environment is amenable to both ontogenetic and phylogenetic levels of analysis. This cross-disciplinary book has been edited to review and integrate the latest research in this area. Various chapters examine which aspects of language (and its foundations) were directly inherited from the common ancestor of humans and nonhuman primates, which aspects have undergone minor change, and which are qualitatively new in *Homo sapiens sapiens*.

The volume has three major themes, woven throughout the chapters. First, it is argued that psychologists and scientists studying animal behaviors, along with researchers in relevant branches of anthropology, need to move beyond unproductive theoretical debate to a more collaborative, empirically focused, and comparative approach to language. Second, accepting this challenge, the contributors describe empirical and comparative methods that reveal some underpinnings of language that are shared by humans and other primates and others that are unique to humans. New insights into the origins of language are discussed, and several hypotheses emerge concerning the evolutionary forces that led to the “design” of language. Third, the volume considers evolutionary challenges (selection pressures) that led to adaptive changes in communication over time with an eye toward understanding the various constraints that channeled this process. Admittedly, this seems a major undertaking (and may even seem preposterous to some), but the investigators involved in this project have the expertise and the data to accomplish it.

Finally, we acknowledge that the writing and publishing of this book was supported by the MEXT grant for the Global COE (Center of Excellence) Research Programme (A06 to Kyoto University).

Nobuo Masataka, Editor

Contents

1 The Gestural Theory of and the Vocal Theory of Language Origins Are Not Incompatible with One Another

Nobuo Masataka

1 Introduction

This book as a whole outlines an approach to the origins of language as the evolution of expressive and communicative behavior of primates, especially until the emergence of single word utterances in *Homo sapiens sapiens* as it is observed currently. It argues that expressive and communicative actions evolved as a complex and cooperative system with other elements of the human's physiology, behavior and social environment.

Even humans, as children, do not produce linguistically meaningful sounds or signs until they are approximately one year old. The ability to produce them begins to develop in early infancy, and important developments in the production of language occur throughout the first year of life. There are a number of earliest major milestones in early interactional development, before the onset of true language, and the accomplishment of most of them requires the children's learning of motor and/or cognitive skills which were inherited by the human species from its evolutionary ancestors. No doubt these skills include both gestural ones and vocal ones. Thus, formulating the question of language origins as either gestural or vocal dichotomously appears irrelevant. Nonetheless scientists concerned with this issue have been preoccupied with determining which of these two hypotheses should be accepted and which should be rejected.

2 Brief History of the Debate about Language Origins

The notion that some animal sounds conveyed semantic information as the human languages did and that iconic visible gestures have something to do with the origin of language is a frequent element in speculation about this phenomenon and appeared early in its history. For example, Socrates hypothesized about the origins of Greek words in Plato's satirical dialogue Cratylus. Socrates's specu-

Primate Research Institute, Kyoto University, Inuyama, Aichi 484-8506, Japan

lation includes a possible role for sound-based iconicity as well as for the kinds of visual gestures employed by the deaf. Plato's use of satire to broach this topic also points to the fine line between the sublime and the ridiculous that has continued to be a hallmark of this sort of speculation (see below).

Such speculation was provided with a somewhat scientific atmosphere when it became joined with the idea that the human species might have a long evolutionary history soon after the publication of Darwin's Origin of Species in 1859. Thereafter there was such an active, one might even use the term rampant, period of speculation that apparently developed into such an annoyance to the Linguistic Society of Paris that it banned the presentation of papers on the subject of the origin of language in 1866. The London Philological Society followed suit in 1872. Thus began a century during which speculation on the origin of language in general fell increasingly into disrepute among serious scholars. However, the historical fact should be noted that just a year before this ban in 1872, Darwin himself published a book called The Descent of Man, in which he devoted some pages to discussing this issue. As detailed in another chapter of mine in this book, he argued that the vocal origin hypothesis is more plausible than the gestural origin hypothesis.

The fact that this book of Darwin became controversial acted as a serious blow to the idea of a gestural origin for language. In 1880, partly as a consequence, at a congress in Milan, the education of the deaf adopted a recommendation that the instruction of deaf students in sign language be discontinued in favor of oral-only instruction. This was not only a watershed event in the education of deaf children, to be followed by a century in which sign languages were suppressed in schools in Europe and the Americas, but it also signaled a general devaluation of and decline in the intellectual status of the history of languages in general and an end to serious scholarly study of the characteristics of language origins.

Historically, we had to wait for the reawakening of serious scientific and scholarly study of the origin of language until the 1970s, when two seminal conferences were held: a symposium at the 1972 meeting of the American Anthropological Association and a subsequent conference hosted by the New York Academy of Sciences in 1975. Apparently, the impetus for this reawakening seems to have been the increasing evidence that could be brought to bear on the subject from paleoanthropology, primatology, neurology, and neurolinguistics (see Christiansen and Kirby 2003 for review).

What is perhaps most evident is that early speculation about language origins following Darwin was severely constrained by a lack of fossil evidence regarding human evolution. At the time of the Paris Society's ban, paleoanthropological knowledge was limited essentially to one skullcap, from the Neander valley (Neanderthal) of Germany, and a few other European fragments, of an extinct relatively recent hominid now thought probably not to have been an ancestor of modern humans. The first finds of the more ancient *Homo erectus* did not come until the 1890s in Java, and those of the still more ancient australopithecines of southern Africa not until the 1920s. Making matters of interpretation more difficult during the first half of the 20th century was the existence of the infamous Piltdown forgery, which presented a picture almost diametrically opposed to that which could be inferred from the erectus and australopithecine material. The forgery was not completely exposed until 1953. Discoveries of fossil humans in Africa, Europe, Asia, and Indonesia have come with increasing frequency in the post World War II era, so that now a fairly coherent story of the course of human anatomical evolution can be pieced together.

During the same post-war period, especially beginning in the 1960s, primatologists from the English-speaking world and Japan were compiling a detailed body of information about the behavior, in the wild and in captivity, of various nonhuman primates, including apes: gorillas, chimpanzees, and gibbons, undoubtedly the closest living relatives of modern *Homo sapiens*, separated from us by what is now known to be a very modest genetic divide. Current attempts to make inferences about the possible language-like behavior of early hominids depend upon a sort of triangulation from the fossil evidence for anatomical characteristics of the various fossil hominids (especially what these might imply about behavior) and what is known about the anatomy and behavior of living nonhuman primates contrasted with the same characteristics of modern humans. Whatever can be inferred through this process of triangulation can be said to be legitimate empirical evidence bearing on the origin and evolution of the human capacity for language prior to the invention of writing, about 5000 years ago. Finally, beginning in the mid-1950s, there was a growing movement to recognize the signed languages of deaf people as bona fide human languages, something that had been generally denied since the late 19th century. Taking together such trends of research in addition to other significant early work on sign language linguistics that began in the early 1970s, Hewes in 1973 proposed that language may have originated in manual gestures rather than in animal calls.

3 Evidence for the Gestural Theory of Language Origins

Since Hewes (1973), scientists supporting this proposal have reported evidence for the notion. Its latest argument is summarized in Corballis' review in the next chapter of this book, in which an evolutionary scenario is documented. What is particularly noteworthy in his argument is, in my view, to understand human speech itself as composed of gestures rather than as elements of discrete sounds. Corballis provides this discussion with recent evidence from articulatory phonology and reaches the conclusion that speech may be part of the mirror system, in which the perception of actions is mapped onto the production of those actions.

This notion is extremely intriguing to me personally as a researcher who has investigated the language learning of preverbal infants. For, even at the very onset of articulated sounds (commonly termed as babbling), any infants, deaf or hearing, are unable to learn to produce them just by hearing alone. Since these units present in babbling are utilized later in natural spoken language, production of babbling of this sort, such as "bababa", "dadada", termed canonical babbling, became taken in 1990s as what marked the entrance of an infant into a developmental stage in which the syllabic foundations of meaningful speech are established. Indeed there is agreement that the onset of canonical babbling is an important developmental precursor to spoken language and that some predictive relations exist between characteristics of the babbling and later speech and language development (see Masataka 2003, for a review).

The empirical evidence has consistently shown that onset of canonical babbling ioccurs in the latter half of the first year in typically-developing infants. Consequently this onset was previously speculated to be a deeply biological phenomenon, geared predominantly by maturation and virtually invulnerable to the effects of auditory experience or other environmental factors (Lenneberg 1967). Such findings reported recently apparently disagree with this argument. A longitudinal investigation revealed that, on the basis of the recording of babbling and other motor milestones in full-term and preterm infants of middle and low socioeconomic status, neither preterm infants whose ages were corrected for gestational age nor infants of low socioeconomic status were delayed in the onset of canonical babbling. That study also showed that hand banging was the only important indicator of a certain kind of readiness to reproduce reduplicated consonant-vowel syllables, and that other motor milestones showed neither delay nor acceleration of onset in the same infants.

Moreover, the onset of repetitive motor action involving the hands is chronologically related to the onset of canonical babbling. We pursued this issue further by conducting meticulous sound spectrographic analyses on all the multisyllabic utterances that were recorded from four infants of Japanese-speaking parents in our longitudinal study. The results of the analyses revealed that the average syllable length of the utterances that did not co-occur with hand banging was significantly longer than that of the utterances that did co-occur with the motor action during the same period. Similarly, the averaged format transition duration of the utterances that did not co-occur with hand banging was significantly longer than that of the utterances that did co-occur with this motor action. These results indicate that some acoustic modifications in multisyllabic utterances take place only when they are co-occurring with rhythmic manual activity. The modifications appear to facilitate infants' acquisition of the ability to produce canonical babbling, because the parameters that were modified when they co-occurred with motor activity concern those that essentially distinguish canonical babbling from earlier speech-like vocalizations. For instance, a vocalization that can be transcribed as /ta/ would be deemed canonical if articulated with a rapid transition duration in a relatively short syllable, but would remain "noncanonical" if articulated slowly. In the latter case, syllables are termed as just "marginal" babbling.

4 Role of Motherese in the Intermediate Stage of Language Evolution

Unless successful with learning to produce canonical babbling, infants are unable to proceed to the following stages of language learning, and failure to produce canonical babbling should eventually result in a considerable delay in reaching those linguistic milestones that are essential for performing various kinds of cognitive learning in general. Such findings apparently constitute evidence for the gestural theory of language origins such as Corballis' hypothesis. Such theories commonly assume that there was a stage in the evolution of language when signs were simply iconic and pantomimic illustrations of the things they referred to. Then, one could imagine a stage during which incidental sounds, especially those that were also iconic or onomatopoetic themselves, came to be associated in a gestural complex with the visible sign and the objects in the world that was being referred to.

Subsequent to this stage, the visible sign could wither away or come to be used as a visual adjunct to the now predominant spoken word. Kita's chapter in this book is an attempt to reconstruct this hypothesized intermediate stage as empirically as possible, focusing his research upon the case of Japanese mimetics. Mazuka and her colleagues are also interested in Japanese mimetics. Cross-linguistically, Japanese language has a relatively such vocabulary. Moreover, many of such vocabulalry items are specifically observed in child-directed speech. Such usage is reported to actually serve as a basis on which young children are helped to learn the language effectively, in terms of its phonology, and is therefore taken to be a sort of "motherese". Their findings, in turn, indicate the existence of the children's perceptual basis for these characteristics of caregiver's speech.

According to the anthropological view (Falk 2004), on the other hand, the evolution of motherese is closely related to the high degree of helplessness in human infants, which is a result of structural constraints that were imposed on the morphology of the birth canal by selection for bipedalism in conjunction with an evolutionary trend for increased brain (and fetal head) size. Thus, unlike the human mother, the chimpanzee mother is able to go about her business with her tiny infant autonomously attached to her abdomen, and with her forelimbs free to forage for food or grasp branches. According to the "putting the baby down" hypothesis, before the invention of baby slings, early bipedal mothers must have spent a good deal of time carrying their helpless infants in their arms and would have routinely freed their hands to forage for food by putting their babies down nearby where they could be kept under close surveillance. Unlike chimpanzee infants, human babies cry excessively as an honest signal of the need for reestablishing physical contact with caregivers, and it is suggested that such crying evolved to compensate for the loss of infant-riding during the evolution of bipedalism. Similarly, unlike chimpanzees, human mothers universally engage in motherese that functions to soothe, calm, and reassure infants, and this, too, probably began evolving when infant-riding was lost and babies were periodically put down so that their mothers could forage nearby. Thus, for both mothers and babies, special vocalizations are hypothesized to have evolved in the wake of selection for bipedalism to compensate for the loss of direct physical contact that was previously achieved by grasping extremities.

In contrast to the relatively silent mother/infant interactions that characterize living chimpanzees (and presumably their ancestors), as human infants develop, motherese provides (among other functions) a scaffold for their eventual acquisition of language. Infant-directed speech varies cross-culturally in subtle ways that are tailored to the specific difficulties inherent in learning particular languages. As a general rule, infants' perception of the prosodic cues of motherese in association with linguistic categories is important for their acquisition of knowledge about phonology, the boundaries between words or phrases in their native languages, and, eventually, syntax. Prosodic cues also prime infants' eventual acquisition of semantics and morphology. The vocalizations with their special signaling properties that first emerged in early hominid mother/infant pairs continued to evolve and eventually formed the prelinguistic substrates from which protolanguage emerged. Therefore, even if language originated as a primarily manual system, its evolution must have occurred, at its very beginning, with the involvement of the auditory system. And once the auditory system was modified, it might have almost inevitably been associated with the modification of the vocal system, by which more effective acoustic transmission of information became possible.

Koda actually presents evidence confirming that possibility in his chapter in this book, reporting the results of detailed acoustic analyses on vocal exchanges of contact calls in free-ranging Japanese macaques. During group progression and foraging, they frequently utter so-called coos to maintain cohesiveness among group members. Usually one animal emits a coo, which is responded to antiphonally by someone. Moreover, unless the spontaneously given coo (designated "the first coo") is replied to, the animal is likely to produce another coo ("the second coo") within a brief interval. Koda made comparative acoustic measurements of such the first and the second coos, and found that when repeated, the second coo became higher in its fundamental-frequency (F0) element and more exaggerated in its frequency modulation, and concluded that these observed modifications should be the rudimentary form of the motherese phenomenon.

5 Implications of Music for Language Evolution

Taken together with the findings described in Yamaguichi and Izumi's, Ghanzafar and Lewkowicz's and Nishimura's chapters, recent studies of macaque coo communication reveal that their vocal behavior is much more flexible than had been assumed previously, and appears somewhat music-like. Moreover, once these characteristics of macaque vocal behavior are recognized as such, it becomes noticeable that the characteristics of interaction between preverbal human infants and their caregivers are also music-like to an almost identical degree. Indeed, we have to wait until the age of 8 months in order to hear truly speech-like vocalizations in infants, and before that time, the manner in which they vocalize closely parallels that in which macaques do, which is summarized in another chapter of my own.

The general consensus about the early interactional development of human infants is that its earliest major milestone is the skill of conversational turn-taking. The ability to participate co-operatively in shared discourse is fundamental to social development in general. When a group of three- to four-month-old infants experienced either contingent conversational turn-taking or random responsiveness in interaction with their Japanese-speaking mothers, contingency was found to alter the temporal parameters of the infant's vocal pattern. Infants tended to produce more bursts or packets of vocalizations when the mother talked to the infant in a random way. When the infants were aged three months, such bursts of vocalization occurred most often at intervals of 0.5–1.5 s, whereas when they were aged 4 months they took place most frequently at significantly longer intervals, of 1.0–2.0 s. This difference corresponded to the difference between intervals with which the mother responded contingently to vocalizations of the infant at the age of three months and four months, respectively. While the intervals (between the onset of the infant's vocalization and the onset of the mother's vocalization) rarely exceeded 0.5 s when the infant was aged three months, they were mostly distributed between 0.5 s and 1.0 s when aged 4 months. After vocalizing spontaneously, the infant tended to pause as if to listen for a possible vocal response from the mother. In the absence of a response, he vocalized repeatedly. The intervals between the two consecutive vocalizations were changed flexibly by the infant according to his recent experience of turn-taking with the mother. Thus, proto-conversational abilities of infants at these ages may already be intentional.

A subsequent series of experiments of mine also demonstrated the fact that, when the adult maintains a give-and-take pattern of vocal interaction, the rate of nonspeech sounds decreases, and instead of such sounds infants produce a greater proportion of speech-like vocalizations. Since the infants are always responded to verbally by the adults, taking turns may facilitate in the infant an attempt to mimic speech-like characteristics of the adult's verbal response. Alternatively, the affective nature of turn-taking could increase positive arousal in the infant, thereby instigating, by contagion, the production of pitch contours contained in the adult's response. On the other hand, it has been shown that if infants receive turn-taking from adults nonverbally, that is, by receiving a nonverbal "tsk, tsk, tsk" sound, this does not affect the speech-like sound/nonspeech sound ratio of the infants.

The timing and quality of adult vocal responses affects the social vocalizations of three-to four-month-old infants. Moreover, once the infant becomes to be framed as a conversational partner, matching starts developing with respect to suprasegmental features of the infant's vocalizations. That is, pitch contours of maternal utterances are likely to be mimicked by the infants. In order to facilitate the infants matching, the caregivers make specific efforts when contingent on with the infants' spontaneous utterances of cooing. When they hear cooing, Japanese-speaking caregivers are more likely to respond nonverbally; they themselves produce cooings in response to the infants' cooing. Moreover, cooing produced by the caregivers is matched with respect to pitch contour with the preceding coo of the infant. Even when the caregivers respond verbally, the pitch pattern of the utterances often imitates that of the preceding infants' cooing (Masataka 2003). Such mimicry is performed by the caregivers without their awareness. Usually they are not conscious of engaging in mimicry. When between three- and four-months old, infants seem not to be aware of the fact that their own vocal production and the following maternal utterance share common acoustic features. However, around the end of the fourth month of life, they acquire the ability to discriminate similarities and differences of pitch contour between their own vocal utterance and the following maternal response. Thereafter, the infants rapidly come to attempt the vocal matching by themselves in response to the preceding utterances of caregivers.

To analyze the developmental processes underlying vocal behavior in infants, a discriminant functional analysis was employed, which statistically distinguishes the infants' cooing following five different types of pitch contours of maternal speech. With this procedure, structural variability in infant vocalizations across variants of maternal speech is found to be characterized by a set of quantifiable physical parameters. The parameters are those that actually distinguish the five different types of maternal speech. Attempts at cross-validation, in which the discriminant profiles derived from one sample of vocalizations are used to classify a second set of vocalizations are totally successful, indicating that the results obtained are not an artifact of using the same data set to derive the profiles and then to test reclassification accuracy. More importantly, the proportion of cross-validated vocalizations that are misclassified decreases as the infant's age increases. Thus, this discriminant analysis is an effective tool to demonstrate that a statistically significant relation develops between the acoustic features of maternal speech and those of the following infant vocalizations as infants grow.

A falling pitch contour is the result of a decrease of subglottal air pressure towards the end of an infant vocalization, with a concomitant reduction in vocal fold tension and length. However, for a rising pitch contour to occur, an increase at the end of the vocalization in subglottal air pressure or vocal fold tension is needed, and thus different, purposeful laryngeal articulations are required. Between the age of four and six months, speech-motor control develops dramatically in infants, associated with changes of the tongue, mouth, jaw and respiratory patterns, to produce vocalizations with distinctively different types of pitch contour. These vocalizations are initially the result of the infants' accidental opening and closing of the mouth while phonating. Six-month-old infants are found to be able to display an obvious contrastive use of different type of pitch contour. The importance of motor learning for early vocal development is greater than has traditionally been assumed (Masataka 1992).

Finally, the problem of which partner is influencing the other is determined experimentally when the controlled prosodic feature of caregiver's vocal behavior is presented to infants. The results show six-month-old infants are able to alter the quality of their responding vocalization according to the quality of preceding maternal speech. Throughout the process of interaction between caregivers and infants it is the caregivers who first become adept at being influenced by what was emitted by the infants on the last turn. Such a behavioral tendency must, in turn, be leaned by the infants. It is on the basis of this learning that the skill of purposeful vocal utterance is considered to be first accomplished by infants.

The purposeful use of one suprasegmental feature of vocalizations, namely pitch contour, plays an important role as a means of signaling different communicative functions before the onset of single words (Halliday 1975). Given this evidence of early use of pitch contour by mothers as a means of interacting, early discrimination and production of pitch contour is the child's first association of language form with respects of meaning. Such early associations may lead the child to later inductions of lexico-grammatical means of cooing similar aspects of meaning. This phenomenon has been investigated in infants exposed to various languages so far. Studies based on naturalistic observations of mother-infant interactions at home, the studies consistently show the association of rising terminal contours with demanding behavior, or protest and of falling contours with "narratives". And it seems to be noteworthy that, around this period, speech-like vocalizations in infancy culminates in the sense that canonical babbling emerges.

6 Musical Origins of Language

Overall, human infants acquire phonology during their first year. However, the newborn has the ability to distinguish virtually all sounds used in all languages at birth in spite of producing no speech sounds. During most of early infancy, music and speech are not as differentiated for very young infants as they are for older children and adults. Early in infancy, caregivers use both speech and music to communicate emotionally on a basic level with their preverbal infants, and it may be that only with experience and cognitive maturation do speech and music become clearly differentiated. As the reason for the occurrence of such a peculiar developmental pattern, we can only note the fact that humans are provided with a finite set of specific behavior patterns, each of which is probably phylogenetically inherited by humans as a primate species. Unlike in nonhuman primates, however, the patterns are uniquely organized during human ontogeny and a coordinated structure emerges that eventually leads us to acquire spoken language. A number of elements can be assembled providing for the onset of language in the infant in a more fluid, task-specific manner determined equally by the maturational status and experiences of the infant and by the current context of the action. Nonetheless, this does not force us to rule out the possibility of either the vocal theory of language origins or the gestural theory of language origins.

References

Christiansen MH, Kirby S (2003) Language evolution. Oxford University Press, Oxford

Darwin CR (1859) On the origin of species. John Murray, London

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

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

Hewes GW (1973) Primate communication and the gestural origin of language. Current Anthropology 14:5–24

Lenneberg EH (1967) Biological foundations of language. Wiley, New York

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

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

2
The Gestural Origins of Language

Michael C. Corballis

1 Introduction

The idea that language evolved from manual gestures dates at least to the philosopher de Condillac (1971/1746), but was revived in modern format by Hewes (1973). The idea was controversial at the time, and remains so, but it continues to be advocated, and appears to have gained increasing acceptance (e.g., Arbib 2005; Armstrong 1999; Armstrong et al. 1995; Corballis 2002; Givòn 1979; Rizzolatti and Arbib 1998; Ruben 2005). From an evolutionary point of view, the idea makes some sense, since nonhuman primates have little if any cortical control over vocalization, but excellent cortical control over the hands and arms. Attempts over the past half-century to teach our closest nonhuman relatives, the great apes, to speak have been strikingly unsuccessful, but relatively good progress has been made toward teaching them to communicate by a form of sign language (Gardner and Gardner 1969), or by using visual symbols on a keyboard (Savage-Rumbaugh et al. 1998). These visual forms of communication scarcely resemble the grammatical language of modern humans, but they are a considerable advance over the paucity of speech sounds that these animals can make. The human equivalents of primate vocalizations are probably emotionally-based sounds like laughing, crying, grunting, or shrieking, rather than words.

Human speech required extensive anatomical modifications, including changes to the vocal tract and to innervation of the tongue, and the development of cortical control over voicing via the pyramidal tract (Ploog 2002). Most of the evidence, discussed in more detail below, suggests that these changes occurred late in hominin evolution, leading some to argue that language itself emerged suddenly, as a "catastrophic" event, with the emergence of our own species, *Homo sapiens*, some 170,000 years ago (Bickerton 1995; Crow 2002). Given the complexity of language, it seems highly unlikely that it could have evolved in all-or-none fashion. A more satisfactory solution, then, is to suppose that grammatical language evolved relatively slowly, perhaps during the Pleistocene, and that the

Department of Psychology, Private Bag 92019, University of Auckland, Auckland 1142, New Zealand

latecomer was not language itself, but rather speech. The gestural theory provides such a solution, since it is likely that the manual system was "language-ready" well before the vocal system was (Arbib 2005).

Although language is often identified with speech, it has become abundantly clear that language can exist independently of speech. Notably, the signed languages of the deaf have all of the essential properties of true language, and are conducted entirely with movements of the hands and face (Armstrong et al. 1995; Emmorey 2002; Neidle et al. 2000). Even in individuals with normal speech, moreover, manual gestures typically accompany speech, and are closely synchronized with it, implying a common source (Goldin-Meadow and McNeill 1999). In many cases, in fact, gestures carry part of the meaning, especially where some iconic reference is needed, as in describing what a *spiral* is (McNeill 1992). Hand and mouth are further linked by the fact that, in most people, the left hemisphere is dominant both for manual action and for vocalization, a coupling often claimed as unique to humans (Corballis 1991; 2003; Crow 2002), even if cerebral asymmetry itself is not (Rogers and Andrew 2002).

2 A Gradual Switch

Nevertheless the gestural theory of language origins has not received widespread acceptance. One of the reasons for this has been succinctly expressed by the linguist Robbins Burling:

> [T]he gestural theory has one nearly fatal flaw. Its sticking point has always been the switch that would have been needed to move from a visual language to an audible one (Burling 2005, p 123).

This argument can be overcome, at least to some extent, if it is proposed that the switch was a gradual one, with facial and vocal elements gradually introduced into a system that was initially primarily manual, although perhaps punctuated by grunts. Through this gradual process, autonomous speech was eventually possible, although even today people characteristically augment their speech with manual gestures (Goldin-Meadow and McNeill 1999).

One argument in favor of a gradual switch has to do with the discovery of the so-called "mirror system" in the primate brain, which underlies manual gesture. In particular, area F5 in the monkey brain includes some neurons, called mirror neurons, that respond both when the animal makes a grasping movement, and when it watches another individual making the same movement. It is now known that area F5 is part of a more general mirror system specialized for the perception of biological motion (Rizzolatti et al. 2001). Area F5 is also thought to be the homolog of Broca's area in the human brain, leading naturally to the suggestion that speech evolved from a primate system involved with manual gestures (Rizzolatti and Arbib 1998).

Discovery of the mirror system bolstered the earlier idea, implied by the motor theory of speech perception (Liberman et al. 1967), that speech itself is fundamentally a gestural system rather than a vocal one. Traditionally, speech has been regarded as made up of discrete elements of sound, called phonemes. It has been known for some time, though, that phonemes do not exist as discrete units in the acoustic signal (Joos 1948), and are not discretely discernible in mechanical recordings of sound, such as a sound spectrograph (Liberman et al. 1967). One reason for this is that the acoustic signals corresponding to individual phonemes vary widely, depending on the contexts in which they are embedded. This has led to the view that they exist only in the minds of speakers and hearers, and the acoustic signal must undergo complex transformation for individual phonemes to be perceived as such. Yet we can perceive speech at remarkably high rates, up to at least 10–15 phonemes per second, which seems at odds with the idea that some complex, context-dependent transformation is necessary.

These problems have led to the alternative view, known as articulatory phonology (Browman and Goldstein 1995), that speech is better understood as comprised of articulatory gestures rather than as patterns of sound. Six articulatory organs—namely, the lips, the velum, the larynx, and the blade, body, and root of the tongue—produce these gestures. Each is controlled separately, so that individual speech units are comprised of different combinations of movements. The distribution of action over these articulators means that the elements overlap in time, which makes possible the high rates of production and perception. Unlike phonemes, speech gestures can be discerned by mechanical means, though X-rays, magnetic resonance imaging, and palatography (Studdert-Kennedy 1998).

The implication is that even the perception of speech is not so much a question of acoustic analysis as one of mapping of speech sounds onto the gestures that produce those sounds, presumably involving an adaptation of the mirror system to include vocalized input. The mirror system is not restricted to visual input even in the monkey brain; Kohler et al. (2002) recorded neurons in area F5 of the monkey that respond to the sounds of actions, such as tearing paper or breaking peanuts. Hence the mirror system was preadapted for the mapping of sounds onto action, but there is no evidence that vocalization is part of the mirror system in nonhuman primates. The evolution of speech, then, involved the incorporation of speech into the mirror system, as part of the more general system for the perception and production of biological motion (Corballis 2003). This probably occurred at some stage after the split between humans and the great apes (Ploog 2002), and possibly only in our own species, as suggested below.

In the course of hominin evolution, it is likely that language increasingly incorporated facial as well as manual movement, especially with the emergence of the use and manufacture of tools. Facial gestures are increasingly recognized as an important component of the signed languages of the deaf. These gestures tend to focus on the mouth, and are distinct from *mouthing*, where the signer silently produces the spoken word simultaneously with the sign that has the same meaning. Mouth gestures have been studied primarily in European signed languages, and schemes for the phonological composition of mouth movements have been proposed for Swedish (Bergman and Wallin 2001), English (Sutton-

Spence and Day 2001) and Italian (Ajello et al. 2001) Sign Languages. Facial gestures also play a prominent role in American Sign Language, providing the equivalent of prosody in speech, and are also critical to many other linguistic functions, such as marking different kinds of questions, or indicating adverbial modifications of verbs (Emmorey 2002). In a recent study Muir and Richardson (2005) found that native signers watching discourse in British Sign Language focused mostly on the face and mouth, and relatively little on the hands or upper body. The face may play a much more prominent role in signed languages than has been hitherto recognized.

The face also plays a role in the perception of normal speech. Although we can understand the radio announcer or the voice on the cellphone, there is abundant evidence that watching people speak can aid understanding of what they are saying. It can even distort it, as in the McGurk effect, in which dubbing sounds onto a mouth that is saying something different alters what the hearer actually hears (McGurk and MacDonald 1976). Evidence from an fMRI study shows that the mirror system is activated when people watch mouth actions, such as biting, lip-smacking, oral movements involved in vocalization, when these are performed by people, but not when they are performed by a monkey or a dog. Actions belonging to the observer's own motor repertoire are mapped onto the observer's motor system, while those that do not belong are not—instead, they are perceived in terms of their visual properties (Buccino et al. 2004). Watching speech movements, and even stills of a mouth making a speech sound, also activates the mirror system, including Broca's area (Calvert and Campbell 2003). This is consistent with the idea that speech may have evolved from visual displays that included movements of the face.

In summary, evidence from spoken and signed language suggests that movements of the hands and face feature prominently in both. This suggests that the evolutionary transition from dominance of the hands to dominance of the face might have been a smooth and continuous one. Vocalization may also have increasingly accompanied gestures of the hands and face, perhaps first in the form of grunts to add emphasis, but gradually incorporating meaning. Even so, vocalization probably did not assume the dominant role until late in hominin evolution, and perhaps only with the emergence of our own species, *Homo sapiens*.

3 The Late Emergence of Vocal Speech

Articulate speech required radical change in the neural control of vocalization. The species-specific and largely involuntary calls of primates depend on an evolutionarily ancient system that originates in the limbic system, but in humans this is augmented by a separate neocortical system operating through the pyramidal tract, and synapsing directly with the brainstem nuclei for the vocal cords and tongue (Ploog 2002). The evidence suggests that voluntary control of vocalization in the chimpanzee is extremely limited, at best (e.g., Goodall 1986). The development of cortical control must surely have occurred gradually, rather than in all-or-none fashion, and perhaps reached its final level of development only in anatomically modern humans. An adaptation unique to *H. sapiens* is neurocranial globularity, defined as the roundness of the cranial vault in the sagittal, coronal, and transverse planes, which is likely to have increased the relative size of the temporal and/or frontal lobes relative to other parts of the brain (Lieberman et al. 2002). These changes may reflect more refined control of articulation and/or more accurate perceptual discrimination of articulated sounds.

Speech also required anatomical changes to the vocal tract. While this too must have been gradual, Lieberman (1998; Lieberman et al. 1972) has argued that the lowering of the larynx, an adaptation that increased the range of speech sounds, was incomplete even in the Neanderthals of 30,000 years ago. Perhaps, then, it was this, rather than the absence of language itself, that kept them separate from *H. sapiens*, leading to their eventual extinction. Lieberman's work remains controversial (e.g., Gibson and Jessee 1999), but there is other evidence that the cranial structure underwent critical changes subsequent to the split between anatomically modern and earlier "archaic" *Homo*, such as the Neanderthals, *Homo heidelbergensis*, and *Homo rhodesiensis*. One such change is the shortening of the sphenoid, the central bone of the cranial base from which the face grows forward, resulting in a flattened face (Lieberman 1998). D. E. Lieberman speculates that this is an adaptation for speech, contributing to the unique proportions of the human vocal tract, in which the horizontal and vertical components are roughly equal in length—a configuration, he argues, that improves the ability to produce acoustically distinct speech sounds.

Also critical to articulate speech was an increase in the innervation of the tongue. The hypoglossal nerve is much larger in humans than in great apes, probably because of the important role of the tongue in speech. Fossil evidence suggests that the size of the hypoglossal canal in early australopithecines, and perhaps in *Homo habilis*, was within the range of that in modern great apes, whereas that of the Neanderthal and early *H. sapiens* skulls contained was well within the modern human range (Kay et al. 1998), although this has been disputed (DeGusta et al. 1999). Changes in the control of breathing were also important for speech, and this is at least partly reflected in the fact that the thoracic region of the spinal cord is larger in humans than in nonhuman primates, probably because breathing during speech involves extra muscles of the thorax and abdomen. Fossil evidence indicates that this enlargement was not present in the early hominids or even in *Homo ergaster*, dating from about 1.6 million years ago, but was present in several Neanderthal fossils (MacLarnon and Hewitt 1999; 2004).

The culmination of changes required for articulate speech may well have occurred very late in the evolution of *Homo*, perhaps even with the arrival of our own species. Some have taken this as evidence that language itself emerged only in *Homo sapiens*. Yet such radical changes must have taken place slowly, over the duration of the Pleistocene at least. This suggests that there must have been a prior form of communication that was shaped in two parallel ways, toward more sophisticated syntax, and both toward a vocal form. There are compelling reasons to suppose that this communication was initially based on manual gestures, but increasingly incorporated movements of the face, and finally articulate vocalization.

4 The *FOXP2* Gene

Genetic evidence confirms the speculation that voicing may have become the dominant characteristic of human language only with the emergence of our own species, *Homo sapiens*. About half of the members of three generations of an extended family in England, known as the KE family, are affected by a disorder of speech and language; the disorder is evident from the affected child's first attempts to speak and persists into adulthood (Vargha-Khadem et al. 1995). The disorder is now known to be due to a point mutation on the *FOXP2* gene (forkhead box P2) on chromosome 7 (Fisher et al. 1998; Lai et al. 2001). For normal speech to be acquired, two functional copies of this gene seem to be necessary.

The nature of the deficit in the affected members of the KE family, and therefore the role of the *FOXP2* gene, have been debated. Some have argued that *FOXP2* gene is involved in the development of morphosyntax (Gopnik 1990), and it has even been identified more broadly as the "grammar gene" (Pinker 1994). Subsequent investigation suggests, however, that the core deficit is one of articulation, with grammatical impairment a secondary outcome (Watkins et al. 2002a). The *FOXP2* gene may therefore play a role in the incorporation of vocal articulation into the mirror system.

This is supported by a study in which fMRI was used to record brain activity in both affected and unaffected members of the KE family while they covertly generated verbs in response to nouns (Liégeois et al. 2003). Whereas unaffected members showed the expected activity concentrated in Broca's area in the left hemisphere, affected members showed relative *under*activation in both Broca's area and its right-hemisphere homologue, as well as in other cortical language areas. They also showed *over*activation bilaterally in regions not associated with language. However, there was bilateral activation in the posterior superior temporal gyrus; the left side of this area overlaps Wernicke's area, important in the comprehension of language. This suggests that affected members may have tried to generate words in terms of their sounds, rather than in terms of articulatory patterns. Their deficits were not attributable to any difficulty with verb generation itself, since affected and unaffected members did not differ in their ability to generate verbs overtly, and the patterns of brain activity were similar to those recorded during covert verb generation. Another study based on structural MRI showed morphological abnormalities in the same areas (Watkins et al. 2002b).

The *FOXP2* gene is highly conserved in mammals, and in humans differs in only three places from that in the mouse. Nevertheless, two of the three changes occurred on the human lineage after the split from the common ancestor with the chimpanzee and bonobo. A recent estimate of the date of the more recent of these mutations suggests that it occurred “since the onset of human population growth, some 10,000 to 100,000 years ago” (Enard et al. 2002, p 871). If this is so, then it might be argued that the final incorporation of vocalization into the mirror system was critical to the emergence of modern human behavior, often dated to the Upper Paleolithic (Corballis 2004).

The idea that the critical mutation of the *FOXP2* gene occurred less than 100,000 years ago is indirectly supported by recent evidence from African click languages. Two of the many groups that make extensive use of click sounds are the Hadzabe and San, who are separated geographically by some 2000 kilometers, and genetic evidence suggests that the most recent common ancestor of these groups goes back to the root of present-day mitochondrial DNA lineages, perhaps as early as 100,000 years ago (Knight et al. 2003). This could mean that clicks were a prevocal way of adding sound to facial gestures, prior to the *FOXP2* mutation.

It is widely recognized that modern humans migrated out of Africa within the past 100,000 years, and eventually spread throughout the globe. The date of this migration is still uncertain. Mellars (2006) suggests that modern humans may have reached Malaysia and the Andaman Islands as early as 60,000 to 65,000 years ago, with migration to Europe and the Near East occurring from western or southern Asia, rather than from Africa as previously thought. This is not inconsistent with an estimate by Oppenheimer (2003) that the eastward migration out of Africa took place around 83,000 years ago. Another recent study suggests that there was back-migration from to Africa at around 40,000 to 45,000 years ago, following dispersal first to Asia and then to the Mediterranean (Olivieri et al. 2006). These dates are consistent with the view that autonomous speech emerged prior to the migration of anatomically modern humans out of Africa. Those who migrated may have already developed autonomous speech, leaving behind African speakers who retained click sounds. The only known non-African click language is Damin, an extinct Australian aboriginal language. *Homo sapiens* may have arrived in Australia as early as 60,000 years ago (Thorne et al. 1999), not long after the migrations out of Africa. This is not to say that the early Australians and Africans did not have full vocal control of speech; rather, click languages may be simply a vestige of earlier languages in which vocalization was not yet part of the mirror system giving rise to autonomous speech.

It is unlikely that the *FOXP2* mutation was the only event in the transition to speech, which undoubtedly went through several steps and involved other genes (Marcus and Fisher 2003). Moreover, the *FOXP2* gene is expressed in the embryonic development of structures other than the brain, including the gut, heart, and lung (Shu et al. 2001). It may have even played a role in the modification of breath control for speech (MacLarnon and Hewitt 1999; 2004). A mutation of the *FOXP2* gene may nevertheless have been the most recent event in the incorporation of vocalization into the mirror system, and thus in the refinement of vocal control to the point that it could carry the primary burden of language.

5 Why Speech?

According to the account presented here, the transition from manual to vocal language was not abrupt. This raises the question, though, of why the transition took place at all. Detailed study of the signed languages of the deaf clearly shows that manual languages can be as sophisticated as vocal ones. Indeed, Emmorey (2005) has argued that if language emerged in the first place as a manual system, it should have remained manual, since there are no compelling reasons to prefer vocal over manual communication. Moreover, one obvious disadvantage of a vocal system is that it involved the lowering of the larynx, which greatly increased the risk of choking to death. Nevertheless, despite Emmorey's arguments to the contrary, there are probably clear advantages to speech over gesture. These advantages are practical rather than linguistic. Clearly, the evolutionary pressure toward speech must have been strong. But why?

There are a number of possible answers. First, a switch to autonomous vocalization would have freed the hands from necessary involvement in communication, allowing increased use of the hands for manufacture and tool use. Indeed vocal language allows people to speak and use tools at the same time, leading perhaps to pedagogy (Corballis 2002). It may explain the so-called "human revolution" (Mellars and Stringer 1989), manifest in the dramatic appearance of more sophisticated tools, bodily ornamentation, art, and perhaps music, dating from some 40,000 years ago in Europe, and maybe earlier in Africa (McBrearty and Brooks 2000). This may well have come about because of the switch to autonomously vocal language, made possible by the FOXP2 mutation (Corballis 2004).

Although manual and vocal language can be considered linguistically equivalent, there are other advantages to vocalization. One factor may have been greater energy requirements associated with gesture; there is anecdotal evidence from those attending courses in sign language that the instructors required regular massages in order to meet the sheer physical demands of sign language expression. The physiological costs of speech, in contrast, are so low as to be nearly unmeasurable (Russell et al. 1998). Further, speech is less attentionally demanding than signed language; one can attend to speech with one's eyes shut, or when watching something else. Speech also allows communication over longer distances, as well as communication at night or when the speaker is not visible to the listener. The San, a modern hunter-gatherer society, are known to talk late at night, sometimes all through the night, to resolve conflict and share knowledge (Konner 1982). A recent study also indicates that the short-term memory span is shorter for American Sign Language than for speech (Boutla et al. 2004), suggesting that voicing may have permitted longer and more complex sentences to be transmitted—although the authors of this study claim that the shorter memory span has no impact on the linguistic skill of signers.

A possible scenario for the switch is that there was selective pressure for the face to become more extensively involved in gestural communication as the hands were increasingly engaged in other activities. Our species had been habitually bipedal from some 6 or 7 million years ago, and from some 2 million years ago was developing tools, which would have increasingly involved the hands. The face had long played a role in visual communication, and as outlined above plays an important role in present-day signed languages. Consequently, there may have been pressure for intentional communication to move to the face, including the mouth and tongue. Gesturing may then have retreated into the mouth, so there may have been pressure to add voicing in order to render movements of the tongue more accessible—through sound rather than sight. In this scenario, speech is simply gesture half swallowed, with voicing added. Even so, lip-reading can be a moderately effective way to recover the speech gestures, and as mentioned earlier the McGurk effect illustrates that speech is in part a visual medium. Adding voicing to the signal could have had the extra benefit of allowing a distinction between voiced and unvoiced phonemes, increasing the range of speech elements.

Arguments for the advantages of speech over sign language are inevitably somewhat post-hoc, and Emmorey (2005) has suggested that some of the arguments mentioned above are unconvincing. Perhaps, though, the proof of the pudding lies in the eating; if sign language is the equal of vocal language, one may ask why it is restricted to the deaf, and is not more widespread. Further, it takes only a slight gain in adaptive fitness for genetic mutations to become fixed in the population: Haldane (1927) computed that a variant resulting in a mere 1% in fitness would increase in population frequency from 0.1% to 99.9% in just over 4,000 generations, a time span that fits easily into hominin evolution, or even into the evolution of our own genus, *Homo*. Changes in the mode of communication can have a dramatic influence on human culture, as illustrated by the invention of writing, and more recently by email and the Internet. These changes were relatively sudden, and cultural rather than biological. The change from manual to vocal communication, in contrast, would have been slow, driven by natural selection and involving biological adaptations, but it may have had no less an impact on human culture—and therefore, perhaps, on human survival.

6 Summary and Conclusions

Since Hewes (1973) presented the case for the gestural origins of language, evidence has accumulated to the point that a plausible scenario can be envisaged. We now have evidence that the adaptations for articulate speech were completed late in hominid evolution, possibly even within the past 100,000 years. Since it is unlikely that language itself evolved so late and so suddenly, this provides a good reason to suppose that grammatical language was previously carried by manual and facial gesture, perhaps with increasing vocal accompaniment. I have suggested also that the final achievement of autonomous speech had a dramatic effect on human culture, and was perhaps even instrumental in the human revolution leading to what has been termed “modern” behavior (Corballis 2004).

The scenario is rendered all the more plausible by the insight that speech itself is a gestural system rather than a vocal one, which is in turn bolstered by the recent discovery of the so-called “mirror system” in the primate brain. Language can thus be conceived as part of the system that directly maps biological action onto perception, present also in primates. Of course language is much more than that, since it involves all of the complexities of grammar, and these features probably emerged well before speech became autonomous. The discoveries about FOXP2 provide further potential insight as to how vocalization might have been incorporated into this system, providing the means by which speech became autonomous. We go from there to telephone and radio—although text messaging may be returning us, or our children, to a visuo-manual mode.

These new developments remain somewhat speculative, but will no doubt add further evidence concerning the evolution of human language. Whether that evidence will continue to support the gestural theory remains to be seen, but that theory has come a long way since Hewes’ formulation in 1973.

References

Ajello R, Mazzoni L, Nicolai F (2001) Linguistic gestures: Mouthing in Italian Sign Languages (LIS). In: Sutton-Spence R, Boyes-Braem P (eds) The hands are the head of the mouth: The mouth as articulator in sign language. Signum-Verlag, Hamburg, pp 231–246

Arbib MA (2005) From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. Behavioral & Brain Sciences 28:105–168

Armstrong DF (1999) Original signs: Gesture, sign, and the source of language. Gallaudet University Press, Washington

Armstrong DF, Stokoe WC, Wilcox SE (1995) Gesture and the nature of language. Cambridge University Press, Cambridge

Bergman B, Wallin L (2001) A preliminary analysis of visual mouth segments in Swedish Sign Language. In: Sutton-Spence R, Boyes-Braem P (eds). The hands are the head of the mouth: The mouth as articulator in sign language. Signum-Verlag, Hamburg, pp 51–68

Bickerton D (1995) Language and human behavior. University of Washington Press, Seattle, WA

Boutla M, Supalla T, Newport EL, Bavelier D (2004) Short-term memory span: Insights from sign language. Nature Neuroscience 7:997–1002

Browman CP, Goldstein LF (1995) Dynamics and articulatory phonology. In: van Gelder T, Port RF (eds) Mind as motion. MIT Press, Cambridge, MA, pp 175–193

Buccino G, Lui F, Canessa N, Patteri I, Lagravinese G, Benuzzi F, Porro CA, Rizzolatti G (2004) Neural circuits involved in the recognition of actions performed by nonconspecifics: An fMRI study. Journal of Cognitive Neuroscience 16:114–126

Burling R (2005) The talking ape. Oxford University Press, New York

Calvert GA, Campbell R (2003) Reading speech from still and moving faces: The neural substrates of visible speech. Journal of Cognitive Neuroscience 15:57–70

de Condillac EB (1971) An essay on the origin of human knowledge. T. Nugent (Tr.). Scholars Facsimiles and Reprints, Gainesville (Originally published 1746)

Corballis MC (1991) The lopsided ape. Oxford University Press, New York

Corballis MC (2002) From hand to mouth: The origins of language. Princeton University Press, Princeton, NJ

Corballis MC (2003) From mouth to hand: Gesture, speech, and the evolution of right-handedness. Behavioral & Brain Sciences 26:199–260

Corballis MC (2004) The origins of modernity: Was autonomous speech the critical factor? Psychological Review 111:543–522

Crow TJ (2002) Sexual selection, timing, and an X-Y homologous gene: Did *Homo sapiens* speciate on the Y chromosome? In: Crow TJ (ed) The speciation of modern *Homo sapiens*. Oxford University Press, Oxford, UK, pp 197–216

DeGusta D, Gilbert WH, Turner SP (1999) Hypoglossal canal size and hominid speech. Proceedings of the National Academy of Sciences 96:1800–1804.

Emmorey K (2002) Language, cognition, and brain: Insights from sign language research. Erlbaum, Hillsdale, NJ

Emmorey K (2005) Sign languages are problematic for a gestural origins theory of language evolution. Behavioral & Brain Sciences 28:130–131

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–871

Fisher SE, Vargha-Khadem F, Watkins KE, Monaco AP, Pembrey ME (1998) Localisation of a gene implicated in a severe speech and language disorder. Nature Genet 18:168–170

Gardner RA, Gardner BT (1969) Teaching sign language to a chimpanzee. Science 165:664–672

Gibson KR, Jessee S (1999) Language evolution and expansions of multiple neurological processing areas. In: King BJ (ed) The origins of language: What nonhuman primates can tell us. School of American Research Press, Santa Fe, NM, pp 189–228

Givòn T (1979) On understanding grammar. Academic Press, New York

Goldin-Meadow S, McNeill D (1999) The role of gesture and mimetic representation in making language the province of speech. In: Corballis MC, Lea SEG (eds) The descent of mind. Oxford University Press, Oxford, pp 155–172

Goodall J (1986) The chimpanzees of Gombe: Patterns of behavior. Harvard University Press, Cambridge, MA

Gopnik M (1990) Feature-blind grammar and dysphasia. Nature 344:715

Haldane JBS (1927) A mathematical theory of natural and artificial selection, part V: Selection and mutation. Proceedings of the Cambridge Philosophical Society 23:838–844

Hewes GW (1973) Primate communication and the gestural origins of language. Current Anthropology 14:5–24

Joos M (1948) Acoustic phonetics. Language Monograph No. 23. Linguistic Society of America, Baltimore, MD.

Kay RF, Cartmill M, Barlow M (1998) The hypoglossal canal and the origin of human vocal behavior. Proceedings of the National Academy of Sciences of the United States of America 95:5417–5419

Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL (2003) African Y chromosome and mtDNA divergence provides insight into the history of click languages. Current Biology 13:464–473

Kohler E, Keysers C, Umilta MA, Fogassi L, Gallese V, Rizzolatti G (2002) Hearing sounds, understanding actions: Action representation in mirror neurons. Science 297:846–848

Konner M (1982) The tangled wing: biological constraints on the human spirit. Harper, New York

Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP (2001) A novel forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413:519–523

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychological Review 74:431–461

Lieberman DE (1998) Sphenoid shortening and the evolution of modern cranial shape. Nature 393:158–162

Lieberman DE, McBratney BM, Krovitz G (2002) The evolution and development of cranial form in Homo sapiens. Proceedings of the National Academy of Sciences of the United States of America 99:1134–1139.

Lieberman P (1998) Eve spoke: Human language and human evolution. W.W. Norton, New York

Lieberman P, Crelin ES, Klatt DH (1972) Phonetic ability and related anatomy of the new-born, adult human, Neanderthal man, and the chimpanzee. American Anthropologist 74:287–307

Liégeois F, Baldeweg T, Connelly A, Gadian DG, Mishkin M, Vargha-Khadem F (2003) Language fMRI abnormalities associated with FOXP2 gene mutation. Nature Neuroscience 6:1230–1237

MacLarnon A, Hewitt G (1999) The evolution of human speech: The role of enhanced breathing control. American Journal of Physical Anthropology 109:341–363

MacLarnon A, Hewitt G (2004) Increased breathing control: Another factor in the evolution of human language. Evolutionary Anthropology 13:181–197

Marcus GF, Fisher SE (2003) FOXP2 in focus: What can genes tell us about speech and language? Trends in Cognitive Science 7:257–262

McBrearty S, Brooks AS (2000) The revolution that wasn't: A new interpretation of the origin of modern human behavior. Journal of Human Evolution 39:453–563

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

McNeill D (1992) Hand and mind: What gestures reveal about thought. University of Chicago Press, Chicago, IL

Mellars P (2006) Going east: New genetic and archaeological perspectives on the modern human colonization of Eurasia. Science 313:796–800

Mellars PA, Stringer CB (eds) (1989) The human revolution: Behavioural and biological perspectives on the origins of modern humans. Edinburgh University Press, Edinburgh

Muir LJ, Richardson IEG (2005) Perception of sign language and its application to visual communications for deaf people. Journal of Deaf Studies & Deaf Education 10:390–401

Neidle C, Kegl J, MacLaughlin D, Bahan B, Lee RG (2000) The syntax of American Sign Language. The MIT Press, Cambridge, MA

Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zaherym N, Scozzari R, Cruciani F, Behar DM, Dugoujon JM, Coudray C, Santachiara-Benerecotti AS, Semino O, Bandelt HJ, Torroni A (2006) The mtDNA legacy of the Levantine Early Upper Palaeolithic in Africa. Science 314:1767–1770

Oppenheimer S (2003) Out of Eden: The peopling of the world. Constable, London

Pinker S (1994) The language instinct. Morrow, New York

Ploog D (2002) Is the neural basis of vocalisation different in non-human primates and *Homo sapiens*? In: Crow TJ (ed) The speciation of modern Homo Sapiens. Oxford University Press, Oxford, UK, pp 121–135

Rizzolatti G, Arbib MA (1998) Language within our grasp. Trends in Cognitive Sciences 21:188–194

Rizzolatti G, Fogassi L, Gallese V (2001) Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews 2:661–670

Rogers LJ, Andrew RJ (2002) Comparative vertebrate lateralization. Cambridge University Press, Cambridge

Ruben RJ (2005) Sign language: Its history and contribution to the understanding of the biological nature of language. Acta Oto-Laryngolica 125:464–467

Russell BA, Cerny FJ, Stathopoulos ET (1998) Effects of varied vocal intensity on ventilation and energy expenditure in women and men. Journal of Speech, Language, and Hearing Research 41:239–248

Savage-Rumbaugh S, Shanker SG, Taylor TJ (1998) Apes, language, and the human mind. Oxford University Press, New York

Shu WG, Yang HH, Zhang LL, Lu MM, Morrisey EE (2001) Characterization of a new subfamily of winged-helix/forkhead (Fox) genes that are expressed in the lung and act as transcriptional repressors. Journal of Biological Chemistry 276:27488–27497

Studdert-Kennedy M (1998) The particulate origins of language generativity: From syllable to gesture. In: Hurford JR, Studdert-Kennedy M, Knight C (eds) Approaches to the evolution of language. Cambridge University Press, Cambridge, UK, pp 169–176

Sutton-Spence R, Day L (2001) Mouthings and mouth gestures in British Sign Language. In: Sutton-Spence R, Boyes-Braem P (eds) The hands are the head of the mouth: The mouth as articulator in sign language. Signum-Verlag, Hamburg, pp 69–85

Thorne A, Grün R, Mortimer G, Spooner NA, Simpson JJ, McCulloch M, Taylor L, Curnoe D (1999) Australia's oldest human remains: Age of the Lake Mungo human skeleton. Journal of Human Evolution 36:591–612

Vargha-Khadem F, Watkins KE, Alcock KJ, Fletcher P, Passingham R (1995) Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. Proceedings of the National Academy of Sciences of the United States of America 92:930–933

Watkins KE, Dronkers NF, Vargha-Khadem F (2002a) Behavioural analysis of an inherited speech and language disorder: Comparison with acquired aphasia. Brain 125:452–464

Watkins KE, Vargha-Khadem F, Ashburner J, Passingham RE, Connelly A, Friston KJ, Frackowiak RSJ, Mishkin M, Gadian DG (2002b) MRI analysis of an inherited speech and language disorder: structural brain abnormalities. Brain 125:465–478

3
World-View of Protolanguage Speakers as Inferred from Semantics of Sound Symbolic Words: A Case of Japanese Mimetics

Sotaro Kita

1 Introduction

The biggest challenge in the study of language evolution is the fact that language does not leave fossil records in the rock. However, it has been argued that some reminants of earlier forms of language has been "fossilized" in modern language. Bickerton (1990) suggested that syntactic properties of "protolanguage" (a communication system that is a precursor to modern language) can be seen in utterances produced by Broca's aphasics, infants in the two-word stage, speakers of a Pidgin language, and Genie, who were deprived of language input until the age of 13 due to abusive imprisonment (Curtis 1977). Jackendoff (2002) suggests that interjections such as *ouch*, *wow*, and *oh* is a fossil from a stage in the development of protolangauge in which words did not combine syntactically and the referents of the words were situation-bound and mostly affective. In this article, we will explore another possible fossil of protolanguage, namely sound-symbolic words. More specifically, this article investigates the semantic properties of these words, taking sound symbolic words in Japanese as an example. Sound symbolic words have certain restrictions as to what type of events and states they can refer to. It is suggested that these restrictions might tell us the "world view" held by the speakers of protolanguage that heavily relied on sound symbolic words.

Sound symbolism has been marginalized in modern linguistics since de Saussure's (1916/1983) influential publication. Saussure stated that one of the most important principles of language is the arbitrary relationship between sound and

University of Birmingham, School of Psychology, Birmingham, B15 2TT, UK

meaning in words. He claims that onomatopoetic words such as *bowwow* (a dog) and *meow* (a cat) are a marginal phenomenon in language: "[Onomatopoetic words] are never organic elements of a linguistic system. Moreover, they are far fewer than in generally believed" (p 69). More recently, Newmeyer (1993) echoes this viewpoint, "the number of pictorial, imitative, or onomatopoetic nonderived words in any language is vanishingly small." (p 758). However, such a statement turns out to be too strong if one looks beyond Indo-European languages.

Many languages of the world have a large word class of sound symbolic words, in which sound and meaning of words are systematically related (Voeltz and Kilian-Hatz 2001). Japanese, for example, has a class of words called mimetics (*giogo*/*gitaigo*). This class of words include onomatopoetic words similar to *bowwow*, but such sound-imitating words constitute only a small part of this word class.

Mimetics can refer not only to sounds but also experiences that are related to vison (*pika*, "a flash of light"; *kira*, "twinkling"), touch (*nurunuru*, "slimy/slippery"; *numenume*, "slimy in a unpleasant way"), taste (*piripiri*, "spicy and hot") and olfaction (*puun*, "stinky"). It can also refer to psychological or physiological states (*sowasowa*, "being nervous"; *kutakuta*, "exhausted"; *zukizuki* "having pulsating pain"). Another domain in which mimetics make fine-grained semantic distinctions is motion. For example, there are over a number of mimetics that refer to different types of human locomotion (Matsumoto 1997): for example, *noshinoshi* "lumber", *yochiyochi* "toddle", *yoroyoro* "shamble", *chokochoko* "walk in small light steps", etc.

As seen in the above examples, mimetics are all semantically very specific. There are no hyponyms or hypernyms among mimetics (Watson 2001). For examples, though there are many mimetics for human locomotion as seen above, there are no hypernym mimetics from them. That is, there are no mimetics that refer to superordinate concepts such as walking and running in the general sense. In turn, there are no hyponyms for existing mimetic, either. Namely, there are no mimetics that further specifies the subtype of a particular locomotion mimetic. This is in contrast to the ordinary non-sound symbolic words (*to walk* is a hypernym for *to shamble*, *to toddle*, *to lumber*, etc., and *to walk* is also a hyponym for *to move*.)

The most important characteristics of mimetics is sound symbolism. Systematic sound-meaning relationships in mimetics can be illustrated by the following examples.

(1)

goro	"a heavy object rolling"
koro	"a light object rolling"
guru	"a heavy object rotating around an axis"
kuru	"a light object rotating around an axis"
bota	"thick/much liquid hitting a solid surface"
pota	"thin/little liquid hitting a solid surface"

The voiced initial consonant tends to indicate a dense, heavy or big object, and the voiceless consonant a light or small object. The velar stops (/k/ and /g/) followed by an /r/ tend to indicate rotation of some kind. Various consonant combinations and vowels are systematically associated with certain meanings (Hamano 1986; 1998).

Mimetics play an important role in the everyday linguistic life of Japanese speakers. Mimetics are open-class words. One mid-sized mimetics dictionary lists 1,700 entries (Atoda and Hoshino 1995). Probably, thousands of words belongs to this category. Mimetics are used frequently in Japanese conversation. They are also used in a wide range of verbal arts: from comic books to novels by Nobel Prize winning authors.

Japanese is not an exception among languages of the world. Many other languages have a large set of open-class words with clear sound symbolism that constitute grammatical classes (Hinton et al. 1994; Nuckolls 1999; Voeltz and Kilian-Hatz 2001). For example, most of sub-Saharan African languages have such a class of words (called "ideophones"; see Childs 1994 for a review). So do many of the southeast Asian languages (called "expressives"; Diffloth 1972; 1976; Watson 2001; Enfield 2005) and some of the Australian Aboriginal languages (Alpher 1994; McGregor 2001; Schultze-Berndt 2001), and indigenous languages in South America (Nuckolls 1996). In Europe, Finish, Estonian (Mikone 2001) and Basque (Ibarretxe-Antuñano 2006) (all non-Indo-European languages) have an extensive sound symbolic word class. In English, systematic sound symbolism has also been described in words such as *squeeze*, *squirt*, *squint*, *squelch*, *bump*, *thump*, *dump*, and *plump* (e.g., Firth 1935/1957), though they do not form a clear grammatical class unlike mimetic words, ideophones and expressives.

From the global perspective, Indo-European languages such as English, German, French are unusual in that they tend to lack a large grammatically defined class of sound symbolic words. Modern linguistics was developed by speakers of these languages. This might explain the emphasis on arbitrariness between form and meaning in words in modern linguistics. Despite being downplayed in modern linguistics, sound symbolism is a robust and wide-spread feature of modern language.

2 What Information Japanese Mimetics Do and Do Not Systematically Encode

Japanese mimetics can refer to a wide range of events and states. Sound symbolism in mimetics systematically encodes various types of information. As we have seen in the previous section, it can encode the mass of moving objects. Interestingly, however, it cannot encode the mass of agents who act upon an object (Kita 1993; 1997), as illustrated in example (2)

(2) (from Kita 1997)
dareka-ga tama-o gorogoro-to korogashi-ta
somebody-NOM ball-ACC Mimetic-COMP roll-PAST
“somebody rolled a heavy ball.”
“somebody heavy rolled a ball.” (Impossible reading)
(See Appendix for the abbreviations used in the gloss.)

In (2), the initial voiced consonant of *gorogoro* “a heavy object rolling” characterizes the mass of the moving object, but it cannot characterize the mass of the person who rolled the ball.

Mimetics systematically encode the temporal structure of an event (Hamano 1998). Reduplication of a mimetic indicates a continuous event, as in (3b).[1]

(3) (Kita 1997, following Hamano 1998)

a.	góro	“a heavy object rolling”
b.	górogoro	“a heavy object rolling continuously”
c.	górogoro górogoro	“a heavy object rolling continuously for a long time”
d.	góro góro	“a heavy object rolling twice”
e.	góro góro góro	“a heavy object rolling three times”

(The accents on the words indicate the locatino of the pitch accent.)

When the reduplicated form (3b) is repeated as in (3c), it indicates a long period of time. When the single form (3a) is repeated, the number of repetition iconically represent the number of repetition in the referent event, as in (3d) and (3e).

Mimetics also systematically distinguish events from states. Mimetics that serves as an adverbial in a sentence express events, in which something changes. In contrast, mimetics that serves as a predicate nominal in a sentence express states, in which there is no change. (Predicate nominals are nouns or noun-like element that serves as a predicate. For example, the noun, *disaster*, in the sentence, *This is a disaster*, is a predicate nominal.) This contrast can be illustrated by a nominal mimetic and an adverbial mimetic that have the same sequence of phonemes, *pikapika*. The adverbial mimetic, *píkapika* “flashing”, has the accent on the first vowel, and the nominal mimetic, *pikápika* “shiney”, has the accent on the second vowel (these are the typical accent placements for adverbial and nominal mimetics). The adverbial mimetic can only have an event reading “flashing”, as in (4a), and the nominal mimetic can only have a state reading, as in (4b).

[1] Reduplication differs from repetition of a word (e.g., (3d)), in that reduplication creates a new word. In order to illustrate this point, the accent in each word is illustrated by an accent diacritic on the vowel. In Japanese, each word has one accent, and the reduplicated form (3b) has one accent, whereas a repeated form as in (3d) has an accent in each of the repeated component.

(4) (Kita 1997)
  a. rampu-ga píkapika-to hikat-te-ir-u
     lamp-NOM Mimetic-COMP glow-COMP-exit-Present
     "The lamp is shiny." (Impossible reading)
     "The lamp is flashing."
  b. rampu-ga pikápika-ni hika-te-ir-u
     lamp-NOM Mimetic-COP glow-COMP-exit-Present
     "The lamp is shiny."
     "The lamp is flashing." (Impossible reading)

Thus, mimetics systematically encode the temporal structure of various experiences. However, other types of time-related information is never encoded in mimetics. For example, the tense information (e.g., past, future, and present in relation to the time of utterance) is never encoded. In other words, there are no mimetics with the meaning like "rolling in the past". Furthermore, mimetics never locate events or state in time such as "at night" or "during the day". In other words, there are no mimetics with the meaning such as "rolling at night" even though ordinary (non-mimetic) words can do so (*yonaki-suru* "to cry at night (as of infants)").

Similarly to the lack of temporal localization of events and states, mimetics do not spatially localize events and states. Thus, there are no mimeitcs with the meaning such as "rolling in the house" even though ordinary (non-mimetic) words can do so (*zaitaku-suru*, "being at one's house").

Finally, mimetics encode affect associated with events and states. For example, the phoneme /e/ adds negative overlay on the referent events and states such as vulgarity (Kindaichi 1978) and inappropriateness (Hamano 1986; 1998), as exemplified by mimetics such as *geragera* "laughing loudly in a vulgar manner" and *dere* "untidy and inappropriate". Kita (1993; 1997) suggested that /e/ in generally overlays the meaning of negative affect towards an event or state, as in (5).

(5)
  a. bita "a wet two-dimentional object sticking"
  b. beta "a wet two-dimentional object sticking, which is unpleasant".

Similar coding of negative affect can be seen in palatalized consonants (Hamano 1986; 1998; Kindaichi 1978). The pairs (6ab) and (6cd) provides phonological minimal pairs in which the mimetics with paratalized consonants such as /hy/ and /ch/ refers to events with negative connotations.

(6) (from Hamano 1986; 1998).
  a. horohoro "noble weeping"
  b. hyorohyoro "being thin and weak"
  c. torotoro "slightly thick liquid moving"
  d. chorochoro "unreliable, unpredicatble movement"

Thus, sound sybmolism in mimetics can encode certain types of information, but not others (see Table 1 for the summary). In the next session, I will discuss what we can infer from such semantic selectivity of mimitics about evolution of language.

Table 1. Types of information that are and are not sound symbolically encoded in mimetics.

Encoded in mimetics

(a) Events and states perceived through all sensory modalities (taste, smell, touch, vision, audition)
(b) Psychological and physiological states (e.g., nervousness, fatigue, pain)
(c) Size, mass, and density of the moving object
(d) Temporal structure of experiences (e.g., event vs states, single event vs. continuous event vs. repeated events)
(e) Affective overlay on events and states (e.g., negative affect)

NOT encoded in mimetics

(a) Size, mass, and density of the agent
(b) Temporal localization of events (e.g., past vs present vs future; at night; in the summer)
(c) Spatial localization of events (e.g., at one's house)

3 Implications for Theories of Language Evolution

The hypothesis that I would like to propose is that the types of information systematically encoded in mimetics reflect the types of cognitive distinctions that were relevant for the "world view" of our ancestors who communicated with protolanguage that heavily relied on sound symbolism. This hypothesis assumes that sound symbolic words in modern language reflect some important aspects of protolanguage.

The possibility of sound symbolic protolanguage has been discussed in the literature, but the discussions have been restricted to onomatopoeias, in which speech sounds imitate sounds in the world. Darwin was sympathetic to the idea that onomatopoeias were an important part of protolanguage. "I cannot doubt that language owes its origin to the imitation and modification, aided by signs and gestures, of various natural sounds, the voices of other animals, and man's own distinctive cries." (Darwin 1871/1955, p 297). In contrast, Müller rejects such a possibility: "no process of natural selection will ever distill significant words out of the notes of birds or the cries of beasts" (Müller 1866, p 354, quoted in Limber 1982). (See more discussions in Johansson 2005). However, these authors did not discuss substantial sound symbolic lexicons in languages like Japanese and others around the world, which go far beyond onomatopoeias. In many languages of the world, sound symbolic words refer to a variety of sensory, psychological, physiological, and affective experiences that are significant in people's lives. Thus, the objection for onomatopoeic protolanguage on the grounds of restricted semantic domains does not apply to sound symbolic protolanguage. In the following section, we first examine evidence for the assumption that sound symbolic words are fossils of protolanguage.

3.1 Sound Symbolic Words as Fossils of Protolanguage

It has been argued that some components of modern language are the vestige of a protolanuage (Bickerton 1990). For example, Jackendoff (2002) argued that interjections such as *ouch* and *oh* are fossils of protolanguage. His idea is supported by the fact that interjections are not fully integrated with the rest of language. For example, interjections do not syntactically combine with other ordinary words to form a complex phrase. A similar argument can be made for Japanese mimetics and other sound symbolic words in the world's languages.

Sound symbolic words are often separated from other words in a sentence in one way or another. In some languages, sound symbolic words appear only at the periphery (beginning or end) of a sentence and not fully syntactically integrated to the sentence (Childs 1994; Diffloth 1976). Japanese mimetics are syntactically integrated with the sentence, and can appear in a mid-sentence position, but it is often used with a quotation complementizer, *to* or *te*.[2] (7) illustrate the use of to with a mimetic (7a) and with a quotation (7b).

(7)

a. tama-ga gorogoro-to korogat-ta.
ball-NOM Mimetic-COMP roll-PAST
"The ball rolled in the *gorogoro*-manner." (see (5) for the meaning of *gorogoro*)

b. Honda-san-ga arigato-to it-ta
Honda-Mr-NOM thank.you-COMP say-PAST
"Mr. Honda said, 'thank you'."

The use of quotation complementizer sets the mimetic apart from the rest of the sentence in the way a quoted utterance is separated from the framing sentence. Note that the use of quotation marker with sound symbolic words is also attested in other languages (e.g., Bartens 2000, p 20; de Jong 2001).

Furthermore, Kita (1993; 1997; 2001) has argued also that sound symbolic words are not fully semantically integrated with the host sentence either. For example, mimetics do not create the sense of wordiness or redundancy even when there is a great referential overlap with the host sentence as in (8a).

(8) (from Kita 1997).

a. Taro-wa sutasuta-to haya-aruki-o si-ta
Taro-TOP Mimetic-COMP haste-walk-ACC do-PAST
"Taro walked hurriedly." (sutasuta = hurried walk of a human)

b.# Taro-wa isogi-ashi-de haya-aruki-o si-ta
Taro-TOP hurried-feet-with haste-walk-ACC do-PAST
"Taro walked hasitily hurriedly."

However, as in (8b), other syntactically optional words such as adverbs would create wordiness or redundancy due to the violation of Griecean Maxim of con-

[2] Mimetics can also appear without the quotation complementizers when they are used as a nominal predicate (in conjunction with a copula verb, *da*) or as a part of a compound verb (in conjunction with a light verb, *suru*, "to do"). It is possible that different types of mimetics lie along the continuum of how well they reflect aspects of a protolanguage. The adverbial mimetics with the quotation complementizers may reflect a protolanguage the best, and the ones used as a nominal predicate may do so the least well.

versation, which states that our utterance should not be longer than necessary. ("#" in (8) indicates pragmatic anomaly of the sentence.)

Another feature of sound symbolic words that set them apart from other ordinary words is that sound symbolic words tend to be phonologically liberal. In other words, sound symbolic words use phonemes or a sequence of phonemes that are not used or rare in the rest of the lexicon. For example, in Japanese mimetics, the phoneme /p/ is commonly used, but this phoneme is not used in the rest of the Japanese lexicon except for loan words such as *pari* "Paris" (McCawley 1968). Sound symbolic words in many other languages are also phonological liberal (e.g., Childs 1994; Ibarretxe-Antuñano 2006; Msimang and Poulos 2001).

These syntactic and phonological features of sound symbolic words indicate that sound symbolic words are not fully linguistically integrated with other ordinary words in the lexicon. This rift between the two types of words can be taken as evidence that sound symbolic words are fossils of protolanguage that have been engulfed and incorporated (albeit not fully) into the system of modern language.

3.2 World-View Expressed in the Sound Symbolic Protolanguage

If sound symbolic words indeed reflect some aspects of protolanguage, then we may be able to infer the property of the protolanguage from sound symbolic words in modern language. We focus on the semantics of sound symbolic words, and aim to infer the "world-view" of the speakers of the sound symbolic protolanguage. The world-view here refers to the way one carves up the world into meaningful elements that are used in communication (Whorf 1940/1956). These semantic distinctions embodied in the world-view are psychologically salient features in the world that are worthy of communication. Thus, from semantics of sound symbolic words, we might be able to infer the thought-world of the speakers of the protolanguage, namely, how the mind of the speakers represented the world.

Let us now turn to the characteristics of the semantics of Japanese mimetics described above (see Table 1). The sound symbolism in mimetics systematically distinguishes various states of affairs that are experienced through different sensory modalities such as vision, audition, etc. or through self-monitoring of psychological, physiological, and affective parameters. It also distinguishes certain internal structures of the state of affairs such as the temporal structuring (e.g., event vs. state, single event vs. repeated event), "manner" of movement (e.g., different ways of walking), and some properties of the moving object (e.g., size, mass). Thus, mimetics can encode a complex set of features of an event, which would normally require several words to describe (see, e.g., (1)).

However, the sound symbolism does not distinguish certain other characteristics that could distinguish events and states from each other. For example, there are no mimetics that spatially localize events and states in relation to other objects or events. In other words, it does not make the distinctions that prepositional phrases (e.g., in English) would make. Similarly, there are no mimetics that temporally localize events and states in relation to other points in time. In other words, it does not make the distinctions that time adverbials and tense would make. This might indicate that mimetics capture moment-by-moment experience of events and states that are not localized in a larger spatio-temporal context. Furthermore, the sound symbolism in mimetics does not distinguish properties of the agent involved as in (7). This contrasts with the fact that mimetics systematically distinguishes properties of the moving object as in (1). In other words, there is no opposition between self and other in terms of agency. This might indicate that mimetics capture subjective experience of events and states in which different agents are not distinguished.

Another interesting feature of the semantics of mimetics is that there are no hyponyms and hypernyms in mimetics. In other words, the events and states that mimetics refer do not have subordinate-superordinate relationships. Mimetics all refer to events and states at the same level of specificity.

Thus, the world-view systematically encoded in mimetics is rich sensory, psychological, physiological, and affective experiences from a subjective viewpoint. Furthermore, these experiences are not anchored in the spatio-temoporal matrix, and not organised into a conceptual hierarchy in terms of subordinate-superordinate relations.

Semantic structure of language can be taken as a reflection of a world-view, consisting of psychologically salient features of the world that are considered to be worthy of communication (Cf., Majid et al. 2004; Sapir 1929; Whorf 1940/1956). Though language and thought cannot be equated, the world-view might provide us with a window into the psychological capacity or orientation of the speakers of the sound symbolic protolanguage. They were able to, first of all, crossmodally map speech sound to information from sensory, psychological, physiological, affective modalities. This is a prerequisite for sound symbolism. Furthermore, they combined information from various sources to form a coherent and fine-grained representation of events and states. However, they did not place their experiences into a larger conceptual matrix. They did not put various concepts into a hierarchy, consisting of superordinate-subordinate relations. Neither did they locate their experiences within temporal and spatial coordinates. In addition, they did not distinguish different agents in their representation of the world. In order to represent different agents, one would have to represent different individuals who have different desires and goals and employ different means to achieve the goals. Thus, the protolanguage speakers might have had very limited abilities to represent the psychological states of other individuals.

3.3 Holistic Protolanguage

In the sound symbolic protolanguage, utterances consisting of a single sound symbolic word were used to express a wide range of experiences. As a single word densely encoded several aspects of an experience such a protolanguage can be characterized as holistic, as opposed to analytic.

The possibility of holistic protolanguage has been discussed in the literature. For example, Wray (2000) proposed that there was a stage of protolanguage that consisted entirely of holistic utterances, in which each utterance is a word that cannot be further analyzed into parts. The inspiration of this hypothesis comes from "formulaic sequences" in modern language such as *by and large*, *happy birthday*, *to pull someone's leg*, which cannot to be analyzed as a sum of components. She proposes that utterances in the holistic protolanguage had arbitrary sound-meaning correspondences, and the function of utterances were interpersonal manipulation and group identity management (e.g., *tebima* for "give that to her", *mupati* for "give that to me", in an imaginary protolanguage).

One of the hallmarks of modern language is its analytic properties (Senghas et al. 2004). Thus, any hypothesis on holistic protolanguage is required to account for how language eventually acquired these properties, i.e., the emergence of syntax. Wray (2000) proposes that analytic decomposition of words, which eventually lead to the emergence of syntax, happened later in the evolutionary history. This decomposition was based on a chance phonetic overlap between two words that share a semantic component. However, such an analytic process that relies on chance overlap has to overcome counterexamples created by the arbitrary association of form and meaning.

Sound symbolic protolanguage, in contrast, offers a more straightforward path to the development of analytic properties of modern language. In sound symbolic protolanguage, words with shared semantic contents share phonetic contents (e.g., see (1)). Thus, the analytic process can easily extract the correlation between subparts of words and meaning components. This leads to analytic decomposition of words, and may eventual engender a linear and hierarchical organisation of words, seen in modern language. Though I am sympathetic to the view that formulaic sequences in modern language reflects some aspects of protolanguage,[3] the sound symbolic protolanguage hypothesis has an advantage in accounting for the evolutionary path beyond holistic utterances.

3.4 Emergence and Development of a Shared Lexicon

Another obvious advantage of sound symbolic protolangauge is that it makes it easier to build an open-ended lexicon shared by a community of speakers (Ramachandran and Hubbard 2001). If a community of speakers share the same biological basis for sound symbolism, then it is easy to reach an agreement as to what labels should be used to which referents. There is evidence that some aspects of sound symbolism are universal, which suggests that it has biological bases. For example, high frequency speech sounds (e.g., /i/ or high tone in a tonal language) are associated with smallness, and low frequency sounds (e.g, /o/ or low tone) are associated with largeness by speakers of different languages (see

[3] At one stage in evolutionary history, precursors of formulaic sequences, sound symbolic words, and interjections may all have been a part of holistic protolanguage.

Ohala 1994 for a review) and even by many animal species (Morton 1994). In addition, voiceless stop consonants (/t/, /k/) are associated with angular shapes, whereas liquids and nasal consonants (/r/, /l/, /m/, /n/) are associated with round shapes (Köhler 1929; Ramachandran and Hubbard 2001). This pattern of associations is found in speakers of different languages (Davis 1961). The cross-linguistic and cross-species validity of sound symbolism support the idea that our ancestors, at some point in evolutionary history, became biologically endowed with sound symbolic capacities, which, in turn, facilitated development of shared lexicons.

4 Conclusions

The previous literature of language evolution has not taken the semantic richness of sound symbolic words into account. This might be due to the fact that Indo-European languages happen not to have a rich sound symbolic system and that scholars of language evolution have mainly been speakers of Indo-European languages. However, a very rich system of sound symbolic words is common in other language families of the world. Most of those words are not onomatopoeias, and refer to non-auditory experiences. We took Japanese as an example (Japanese sound symbolic words are called "mimetics"), and discussed what we might be able to infer about protolanguage and its speakers.

It was argued that sound symbolic words are fossils of protolanguage as those words are aberrant in terms of liberal phonology and weak syntactic and semantic integration with the host sentence. Sound symbolic words refer to sensory, psychological, physiological, and affective experiences. The sensory experiences include those based on vision, touch, taste, and smell, as well as those based on audition. Thus, sound symbolic words, and by inference, sound symbolic protolanguage, can refer to a wide range of experiences. However, we can infer semantic limitations of sound symbolic protolanguage, based on the limitations in sound symbolic words in modern language. In sound symbolic words, the referent concepts are not placed in relation to other concepts. Sound symbolic words do not systematically encode the time and location of the referent event or state. They also do not have hypernym-hyponym relations with each other. In other words, all words refer to aspects of the world at the same degree of specificity. Furthermore, sound symbolic words do not systematically encode properties of agents, though they encode properties of moving objects (e.g., size, mass, and density).

It was also argued that these characteristics of sound symbolic words may reflect the world-view (psychologically salient aspects of the world) of the speakers of sound symbolic protolanguage. These ancestors may be able to combine information from sensory, physiological, and emotional sources to form a unified sense of experience. However, they may not relate these experiences in a larger conceptual matrix with spatio-temporal coordinates and hierarchical organizations. They may also not distinguish different agents, who have different desires and goals and may chose different means to achieve the goals. In other words, they might have had a limited capacity to represent other individuals' mental states.

Finally, the sound symbolic protolanguage hypothesis offers a straightforward explanation for two important steps in language evolution. First, it explains how a community of speakers could develop a shared open-ended lexicon. Once the biological basis for sound symbolism emerged in our ancestors, they could use sound symbolism to kick-start and further develop a shared lexicon. Second, sound symbolic protolanguage would be a good stepping stone to analytic protolanguage with word combinations. In sound symbolic words, a semantic overlap entails a phonetic overlap. And, these overlaps could be used to analyze a once unanalyzed holistic word into components, which eventually might have lead to the emergence of syntax.

References

Alpher B (1994) Yir-Yoront ideophones. In: Hinton L, Nichols J, Ohala JJ (eds.) Sound symbolism. Cambridge University Press, Cambridge, pp 161–177

Atoda T, Hoshino K (1995) Giongo gitaigo tsukaikata jiten (Usage dictionary of sound/manner mimetics). Sotakusha, Tokyo

Bartens A (2000) Ideophones and sound symbolism in Atlantic creoles. Academia Scientiarum Fennica, Helsinki

Bickerton D (1990) Language & species. University of Chicago Press, Chicago IL

Childs GT (1994) African ideophones. In: Hinton L, Nichols J, Ohala JJ (eds) Sound symbolism. Cambridge University Press, Cambridge, pp 178–206

Curtis S (1977) Genie: A linguistic study of a modern-day "Wild child". Academic Press, New York

Darwin C (1955) The descent of man and selection in relation to sex. Encyclopaedia Britannica, Chicago (The original work published 1871)

Davis R (1961) The fitness of names to drawings: A cross-cultural study in Tanganyika. British Journal of Psychology 52:259–268

Diffloth G (1972) The notes on expressive meaning. In: Peranteau PM, Levi JN, Phares GC (eds) Papers from the eighth regional meeting of the Chicago linguistic society. Chicago Linguistic Society, Chicago, IL, pp 440–447

Diffloth G (1976) Expressives in Semai, *Austroasiatic Studies* 1:249–264

Enfield NJ (2005) Areal linguistics and mainland Southeast Asia. Annual Review of Anthropology 34:181–206

Firth JR (1935/1957) The use and distribution of certain English sounds. In: Firth JR (ed) Papers in linguistics 1934–1951. Oxford University Press, London, pp 34–46 (Reprinted from Firth JR (1935), English Studies 17:2–12)

Hamano S (1986) The sound-symbolic system of Japanese. Ph.D. dissertation, University of Florida

Hamano S (1998) The sound symbolic system of Japanese. CSLI Publications & Kuroshio, Stanford, CA & Tokyo

Hinton L, Nichols J, Ohala JJ (1994) Sound symbolism. Cambridge University Press, Cambridge

Ibarretxe-Antuñano I (2006) Sound symbolism and motion in Basque. Lincom, München

Jackendoff R (2002) Foundations of language: Brain, meaning, grammar, evolution. Oxford University Press, Oxford

Johansson S (2005) Origins of language: Constraints on hypotheses. John Benjamins, Philadelphia, PA

de Jong N (2001) The ideophone in didinga. In: Voeltz FKE, Kilian-Hatz C (eds) Ideophones. John Benjamins, Amsterdam, pp 121–138

Kindaichi H (1978) Giongo ginatigo gaisetsu (An introduction to mimetics). In: Asano T (ed) Giongo gitaigo jiten (Mimetics dictionary) Kadokawa, Tokyo, pp 3–25

Kita S (1993) Language and thought interface: A study of spontaneous gestures and Japanese mimetics. University of Chicago, Chicago, IL

Kita S (1997) Two-dimensional semantic analysis of Japanese mimetics. Linguistics 35:379–415

Kita S (2001) Semantic schism and interpretive integration in Japanese sentences with a mimetic: A reply to Tsujimura. Linguistics 39:419–436

Köhler W (1929) Gestalt psychology. Liveright Publishing Corporation, New York

Limber J (1982) What can chimps tell us about the origin of language? In: Kuczaj S (ed) Language development, vol 2. Lawrence Erlbaum, Hillsdale, NJ, pp 429–469

Majid A, Bowerman M, Kita S, Haun DB, Levinson SCL (2004) Can language restructure cognition? The case of space. Trends in Cognitive Sciences 8:108–114

Matsumoto Y (1997) Kuukan idoo no gengohyoogen to sono kakuchoo (Linguistic expression of motion in space, and their extensions). In: Tanaka S, Matsumoto Y (eds) Kuukan to idoo no hyoogen (Expressions of space and motion). Kenkyuusha, Tokyo, pp 125–230

McCawley JD (1968) The phonological component of a grammar of Japanese. Mouton, The Hague

McGregor W (2001) Ideophones as the source of verbs in northern Australian languages. In: Voeltz FKE, Kilian-Hatz C (eds) Ideophones. John Benjamins, Amsterdam, pp 205–221

Mikone E (2001) Ideophones in Balto-Finnic languages. In: Voeltz FKE, Kilian-Hatz C (eds) Ideophones. John Benjamins, Amsterdam, pp 223–234

Morton ES (1994) The biological bases of sound symbolism. In: Hinton L, Nichols J, Ohala JJ (eds) Sound symbolism. Cambridge University Press, Cambridge, pp 348–365

Msimang CT, Poulos G (2001) The ideophone in Zulu: A re-examination of conceptual and descriptive notions. In: Voeltz FKE, Kilian-Hatz C (eds) Ideophones. John Benjamins, Amsterdam, pp 235–249

Müller FM (1866) Lectures on the science of language. Scribner, New York

Newmeyer FJ (1993) Iconicity and generative grammar. Language 68:756–796

Nuckolls JB (1996) Sounds like life. Oxford University Press, New York

Nuckolls JB (1999) The case for sound symbolism. Annual Review of Anthropology 28:225–252

Ohala JJ (1994) The frequency code underlies the sound-symbolic use of voice pitch. In: Hinton L, Nichols J, Ohala JJ (eds) Sound symbolism. Cambridge University Press, Cambridge, pp 325–347

Ramachandran VS, Hubbard EM (2001) Synaesthesia—A window into perception, thought, and language. Journal of Consciousness Studies 8:3–34

Sapir E (1929) The status of linguistics as science. Language 5:207–214

de Saussure F (1983) Course in general linguistics, R. Harris, trans. Open Court, La Salle, IL (Original work published 1916)

Schultze-Berndt E (2001) Ideophone-like characteristics of uninflected predicates in Jaminjung (Australia). In: Voeltz FKE, Kilian-Hatz C (eds) Ideophones. John Benjamins, Amsterdam, pp 355–374

Senghas A, Kita S, Özyürek A (2004) Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. Science 305:1779–1782

Voeltz FKE, Kilian-Hatz C (eds) (2001) Ideophones. John Benjamins, Amsterdam

Watson RL (2001) A comparison of some Southeast Asian ideophones with some African ideophones. In: Voeltz FKE, Kilian-Hatz C (eds) Ideophones. John Benjamins, Amsterdam, pp 385–405

Whorf BL (1956) Science and linguistics. In: Carroll JB (ed) Language thought and reality. The MIT Press, Cambridge, MA, pp 207–219 (The original work published 1940)

Wray A (2000) Holistic utterances in protolanguage: The link from primates to humans. In: Knight C, Studdert-Kennedy M, Hurford J (eds) Evolutionary emergence of language: Social function and the origins of linguistic form. Cambridge University Press, West Nyack, NY, pp 285–302

Appendix

Abbreviations used in glossing Japanese examples

ACC	accusative
COMP	complementizer
COP	copula
TOP	topic

4
Japanese Mothers' Use of Specialized Vocabulary in Infant-Directed Speech: Infant-Directed Vocabulary in Japanese

Reiko Mazuka[1,2], Tadahisa Kondo[3], and Akiko Hayashi[4]

1 Introduction

1.1 Infant-Directed Vocabulary

When adults talk to infants or young children, they modify their speech. The specialized speech is sometimes called "motherese" or "infant-directed speech" (IDS). Many characteristics of IDS have been documented across many languages, but the best known characteristics of IDS have to do with prosody of the speech, such as higher pitch and exaggerated pitch contours, and longer more frequent pauses (c.f., Fernald and Simon 1984; Fernald and Kuhl 1987; Fernald and Mazzie 1991; Snow and Ferguson 1977). Other types of modifications also occur, such as changes in syntactic properties, e.g., shorter and simpler utterances, and semantic contents, e.g., conversation about "here and now" (c.f., Newport et al. 1977). It has often been argued that many of the IDS properties are universal (Fernald 1993; Fisher and Tokura 1996; Grieser and Kuhl 1988; Kuhl and et al. 1997; Trainer and et al. 2000), but there are significant cross-linguistic variations in the way mothers interact with their infants (e.g., Fernald and Morikawa 1993), and the way adults modify their speech in IDS (Fernald et al. 1989).

Japanese infant-directed speech is often characterized by a frequent use of vocabulary that is non-adult like (Murase et al. 1992; Murata 1960; 1968; Ogura et al. 1993). A large number of onomatopoetic expressions and reduplications are used instead of the regular adult form, resulting in specialized vocabulary that is often unrelated to the adult form phonologically. For example, "gohan" (food) becomes "maNma", and "ashi" (foot) becomes "aNyo." Specialized vocabulary of this type can be found in other languages as well. In English, for example, words such as "pee-pee" and "poh-poh" are used only with small children, and words

[1] Laboratory for Language Development, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan, [2] Duke University, Department of Psychology and Neuroscience, Box 90086, Durham, NC 27708-0086, USA, [3] NTT Communication Science Laboratories, 3-1 Morinosato, Wakamiya, Atsugi, Kanagawa 243-0198, Japan, and [4] Tokyo Gakugei University, 4-1-1 Nukuikita-machi, Koganei, Tokyo 102-8160, Japan

such as “mommy” and “doggy” instead of “mom” and “dog” may also be used with children primarily. But the use of specialized vocabulary in Japanese is particularly extensive. Murata (1960) surveyed 100 mothers with one-year-old infants, and found that of the 378 basic vocabulary items included in his survey, many mothers reported that they use the specialized form for approximately half of the items (as opposed to the adult forms) when talking to their children. As a testimony of the wide-spread use of this type of vocabulary, there is a published dictionary of child-directed vocabulary, detailing regional variation of the child-directed word forms (“Zenkoku Yoojigo Jiten” (National Child-vocabulary Dictionary), Tomonaga 1997). These types of words are often referred to as “*ikujigo*” in Japanese, which literally means “child-rearing words.” In this paper, we will refer to the specific form of vocabulary Japanese adults would use with infants and young children as “infant-directed vocabulary” (IDV).

1.2 Mora in Japanese Phonology

In this paper, we will analyze the phonological and prosodic characteristics of IDV. As a background, we describe the basic prosodic properties of Japanese here. Japanese is often characterized as having a “mora-timed” rhythm (c.f., Abercrombie 1967; Ladefoged 1975; Port et al. 1987). Mora is a sub-syllabic unit that determines the weight of a syllable, as shown in (1).

Foot (Ft)
|
Syllable (σ) (1)
|
Mora (μ)

Japanese allows three types of syllables; V, CV and CVC. The majority of syllables are mono-moraic and they are either V or CV with a short vowel. Coda consonants are limited to either moraic nasal /N/ (e.g., “n” of /hoNda/ Honda), or the first half of geminate obstruents /Q/ (e.g., first “s” of /niQsaN/ Nissan). These two types of coda consonants and the second half of long vowels /H/ and diphthongs /J/ count as one mora, and they are sometimes called “special morae”. Japanese syllables that contain one of the special morae are thus bimoraic.

The mora is considered to be the unit for Japanese rhythm. So a Japanese speaker tends to produce each mora at a regular rate (but see, Warner and Arai 2001, for review). In addition to being the rhythm-bearing units, morae also play important roles in other ways in Japanese phonology (Kubozono 1995; 1999). It is also argued that Japanese adults use morae as a unit of word segmentation (Otake 2006; Otake et al. 1993; 1996).

We conducted a survey of Japanese IDV by Japanese mothers with young infants. In order to examine the significance of IDV characteristics, we analyzed a number of other types of Japanese samples and compared them to the IDV results: an additional survey data from young women without children, a corpus of infant-directed speech in Japanese (Mazuka et al. 2006), a published corpus of adult spoken Japanese (Maekawa 2006), a database of Japanese word familiarity (Amano and Kondo 1999), and Japanese newspaper articles (Amano and Kondo 2000).

Throughout this paper, we will describe our stimuli in terms of R(egular) mora and S(pecial) mora. As needed, we will also describe them in terms of H(eavy) and L(ight) syllables. The reasons for this are two folds: First, Japanese orthography is mora-based (i.e., *kana* characters), and most descriptions of Japanese lexical items, including dictionary listings, provide *kana* transcription for each word. As the *kana* transcriptions for long vowels and diphthongs are not transparent, syllable weight cannot be computed accurately without additional coding for long vowels and diphthongs. By using the mora-based description, we are able to analyze a large quantity of corpus and dictionary data without compromising accuracy. Second, as discussed above, Japanese speakers' intuitions about word-length are mora-based. For example, three-mora words such as /riNgo/ (apple) and /banana/(banana) are judged to be the same length irrespective of their number of syllables, while two-mora words such as /kaki/(persimmon) are judged to be shorter than either /riNgo/ or /banana/. Thus, the use of a mora-based description, and comparison of syllable weights among the words with the same mora-length, reflects Japanese native speakers' intuitive judgments about word-length. The use of mora-based description, however, is not intended to imply that Japanese children's representation of lexical items are mora-based.

2 Data

2.1 Japanese Infant-Directed Vocabulary Survey

It has been well established that infants prefer to listen to infant-directed speech as opposed to adult-directed speech from a very young age, and it has often been argued that it is due to the exaggerated prosody of infant-directed speech (c.f., Cooper and Aslin 1990; Fernald and Kuhl 1987; Hayashi et al. 2001). If the specialized form of infant-directed vocabulary in Japanese plays a significant role for Japanese learning infants, it is likely that it is the prosodic property of IDV that is important. Thus, we focused on the prosodic form that typical infant-directed vocabulary takes in Japanese.

As discussed above, there exists substantial research that investigated various aspects of Japanese infant-directed vocabulary (IDV). However, we do not have access to original raw data and cannot analyze the prosodic forms of infant-directed vocabulary from previous studies. Thus, we conducted a survey of Japanese mothers with infants about their intuitive judgment of infant-directed vocabulary.

2.1.1 Method

The primary data for our study is a survey we conducted with Japanese mothers. 23 Japanese mothers from the Tokyo area with 8 to 12 month-old infants responded to the survey by mail. The mothers had previously participated in an infant speech perception experiment at Tokyo Gakugei University. The mothers were between 23 and 37 years of age (average age = 29.9), and 12 of them had girls and 11 of them had boys. Four of them also had an older child. 16 additional mothers were sent the survey, but did not respond. They were told “When adults talk to children, they sometimes use special kinds of words, such as ‘otete’ (hand) and ‘neNne’ (sleep). Think of words you would use with young children. List as many as you can.”

To evaluate whether or not the infant-directed vocabulary gathered from mothers specifically comes from them being mothers and interacting with their children, we also gave the same survey to young women who are not mothers. 10 Japanese women, between 20 and 25 years of age, who are single, have no children, and have had little experience interacting with young children answered the same survey.[1]

2.1.2 Results of IDV Survey

In the mothers’ survey, a total of 509 words, with an average of 22 words per mother (range 7–54 words) were reported. Many of the same words were reported by multiple mothers. When we counted the occurrence of different words, there were 237 wordtypes. From non-mothers, a total of 245 words, with an average of 24 words per woman (range 10–58 words) were reported. 137 of them were different items. For each group, we coded each word for word length in morae, the presence or absence of special morae and the position of special morae when they occur. For our analysis, the second vowels of two vowel sequences such as “ai”, “au”, “ae”, “ui” and “oi” that would be typically pronounced as diphthongs are coded as special morae. From this, we classified the words in terms of their mora and syllable length, and calculated how often each type of words was reported for each participant.

2.1.2.1 Prosodic Form of IDV

As shown in Table 1, the most frequent IDV words were 3- and 4-mora words in both groups. In terms of syllables, 2-syllable words were by far the most frequent words. We then analyzed the syllable weight of 3- and 4-mora words. Figure 1 shows the proportion of different word types among 3-mora and 4-mora words. As can be seen from this figure, a large portion of the IDV was in two prosodic forms. The largest number of IDV words were 3-mora, 2-syllable long, and they were of the form RSR (R = regular mora, S = special mora). In terms of syllables, they are heavy-light (HL) bisyllabic words. On average, 79% of the 3-mora IDV words reported by the mothers and 80% by the non-mothers were of this form. In contrast, only 18% of 3-mora words were of the RRR (LLL) type, as reported by mothers and by non-mothers, which is less than 1/4 of the RSR type.

[1] The exact ages of the individual participants were not available.

Table 1. Number of IDV words reported by mothers and non-mothers in terms of mora and syllable length.

	Length in morae								
Length in syllables	1	2	3	4	5	6	7	8	Total
Mothers									
1		18	6						24
2		15	170	133					318
3			36	9	13	4			62
4				55	6	17	3	4	85
5						4	2	1	7
6						6	1	4	11
7									0
8								2	2
Total	0	33	212	197	19	31	6	11	509
Non-mothers									
1	1	9							10
2		8	93	55					156
3			21	5		1			27
4				40	1	1		1	43
5						1	4		5
6						1	1	1	3
7								1	1
8									0
Total	1	17	114	100	1	4	5	3	245

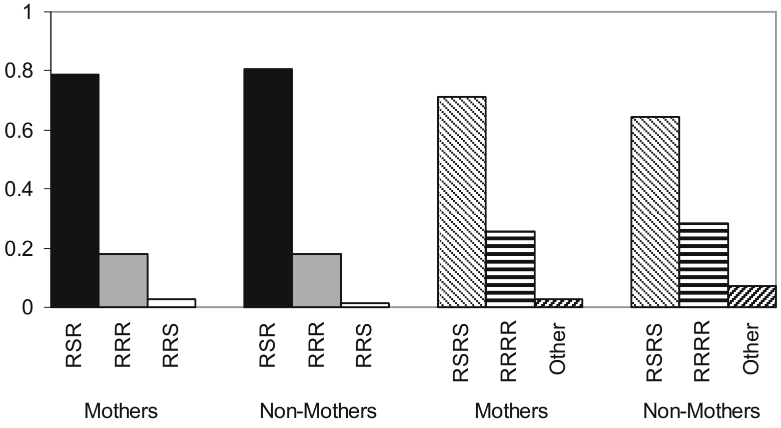


Fig. 1. IDV Survey: proportion of word types for mothers and non-mothers among 3- and 4-mora words.

TABLE 2. Frequently reported items. Items reported by more than half of participants. HL = RSR, HH = RSRS.

Mothers			
Item in Japanese	Translation	Word type	# Of mothers
maNma	Food	HL	22
waNwaN	Dog	HH	18
aNyo	Leg/foot/walk	HL	17
buRbuR	Car	HH	17
nyaNnyaN	Cat	HH	17
neNne	Sleep	HL	15
taQchi	Stand	HL	12

Non-mothers			
Item in Japanese	Translation	Word type	# Of non-mothers
maNma	Food	HL	9
aNyo	Leg/foot/walk	HL	8
waNwaN	Dog	HH	8
buRbuR	Car	HH	7
neNne	Sleep	HL	7
buRbu	Car	HL	6
kuQku	Shoe	HL	6
otete	Hand	LLL	6

4-mora, 2-syllable RSRS words were the second most frequent words, which are heavy-heavy (HH) words. On average, 71% of 4-mora IDV words reported by mothers and 64% by non-mothers were of this type. RRRR words (LLLL), which are the second most frequent 4-mora words occurred 26% of the time in mothers' 4-mora words, and 28% of the time in non-mothers. The two most frequent word types accounted for 62% and 65% of the entire IDVs reported by the mothers and non-mothers respectively. As shown in Table 2, almost all of the frequently reported IDV by the mothers and non-mothers were of either of these two types.

2.1.2.2 Onomatopoeia and Mimetics

Many of the IDV words contained some form of repetition. Among the RSR type words, the most frequent pattern is the /maNma/ type, where the first mora of the heavy syllable is repeated in the second syllable. Another type involves adding the prefix /o/ that indicates politeness to a word, and repeating either a mora or a bimoraic unit. For example, the mono-moraic word /te/ means hand in adult Japanese. This becomes /otete/ in IDV, by adding /o/ at the beginning and repeating /te/. Among 4-mora words, repetition of two morae as a unit was by far the most frequent pattern. If the word is of the RSRS type, the heavy syllable is reduplicated, e.g., /waNwaN/ (dog). In RRRR words, two light syllables are repeated, e.g., /gorogoro/ (lie down). Many of these bimoraic repetitions come from onomatopoetic or mimetic expressions, which are mostly 4 morae long. But there were many words with repetition that cannot be considered onomatopoetic or mimetic e.g., /kamikami/ (chewing; a conjunctive form of the verb “kamu” (chew) repeated), /nainai/ (put away; the adjective “nai” (non-present) repeated). In fact, 65% of the IDV reported by mothers, and 64% by non-mothers contained some form of repetition. In contrast, only 39% of the IDV reported by mothers and 38% by non-mothers were onomatopoeia or mimetic. This shows that the repetition of a syllable or bimoraic unit is a major characteristic of Japanese IDV, but this property cannot be attributed entirely to the fact that many of the IDV words are derived from onomatopoetic or mimetic expressions. In the following sections, we will call the two dominant word forms of the IDV survey (RSR and RSRS words) IDV forms.

2.1.2.3 Mothers vs Non-mothers

The results of the IDV survey from mothers and non-mothers were analyzed in the same way. The comparison of the two groups shows that the responses given by mothers and non-mothers are extremely similar in all of the measures we analyzed.

One-way ANOVAs between the two groups showed that in none of the following measures, the two groups differed significantly—number of items reported by each participant, length of words in terms of mora ($F_{1,31} = 1.39$; $P = 0.25$) or syllable, proportion of RSR words among 3-mora words, RSRS words among 4-mora words, frequency of onomatopoetic and mimetic expressions, frequency of repetition ($F_{1,31} < 1$ in all comparisons except otherwise mentioned).

2.2 Spoken Corpora

The survey of IDV reflects mothers’ intuitive judgment of what IDV is like. It does not necessarily mean that the frequently reported IDV words actually occur frequently in infant-directed speech. To compare the actual occurrence of IDV in infant-directed speech, we analyzed the actual occurrence of word-like units in a Japanese infant-directed speech corpus (RIKEN Japanese Mother-Infant Conversation Corpus R-JMICC, Igarashi and Mazuka 2006; Mazuka et al. 2006).

Another important question is whether the frequent occurrence of RSR and RSRS type words is a characteristic of infant-directed speech, or a reflection of adult Japanese speech characteristics. In English, trochees, i.e., the strong-weak sequence of syllables, are the predominant pattern, and this pattern occurs much more frequently than the iambic pattern (c.f., Cutler and Carter 1987). In child-directed speech, this asymmetry becomes more exaggerated (Kelly and Martin 1994). Thus, it is possible that the frequent occurrence of the IDV forms is an exaggeration of adult Japanese speech. To address this issue, we analyzed a number of adult Japanese speech data and compared them to IDV survey results. The analysis of the R-JMICC corpus allows us to compare spoken corpora of ID and AD within the same speakers. As an example of very different styles of spoken corpus, we also analyzed a published corpus of spoken Japanese (Corpus of Spontaneous Japanese, Maekawa 2006).

2.2.1 Corpus of Infant-Directed and Adult-Directed Speech

R-JMICC is a corpus of 22 Japanese mothers' conversations with their 18 to 24 month-old infants and with an adult experimenter. The mothers were between 25 and 43 years of age (average age 33 years 8 months). 11 of the infants were the only child of the mothers in the corpus, 10 of them had one older sibling, and one had one older and one younger sibling. For each mother-infant dyad, recordings were made for approximately half an hour for mother-infant conversation in book reading and free play, and approximately 15 minutes for adult-adult conversation between the mother and an experimenter on topics related to childrearing (the experimenter was a 33-year-old female whose own infant was 23 months old at the time of recording). The total duration of ID speech was approximately 11 hours, with 50,000 words, while the total of AD speech was about 3 hours long, with 30,000 words.

The corpus is morphologically coded for short-unit words (SUW) and long-unit words (LUW), based on the definition developed for Corpus of Spontaneous Japanese (CSJ; Maekawa 2006). SUW are mono-morphemic words or words made up of at most two morphemes that roughly correspond to dictionary entries of Japanese dictionary. LUWs are combinations of words that sometimes correspond to compound words. For example, /baikiNmaN/ (a Japanese children's cartoon character, "Germs-man") may be analyzed as two SUWs /baikiN/(germs) and /maN/(man), while it can also be analyzed as a single LUW. The entire corpus is also labeled for phonetic segments. For this purpose, phoneme-based labels were used with additional coding for relevant phonetic details, such as vowel devoicing or palatalization of consonants before /i/. Non-syllabic morae, such as the second half of a long vowel /H/, moraic nasal /N/ and the second part of a geminate obstruent /Q/, were also labeled. Non-linguistic usage of similar phonetic events, such as non-lexical lengthening of vowels, <H> and non-lexical lengthening of consonants, <Q> are also coded, making it possible to analyze phenomena such as lexical and non-lexical vowel lengthening independently.

2.2.2 Corpus of Spontaneous Japanese

As a sample of a very different type of spoken corpus, we analyzed data from the Corpus of Spontaneous Japanese (Maekawa 2006), developed by The National Institute for Japanese Language. It consists of a total of 650 hours of public speeches, such as academic conference presentations and simulated speeches, given to small friendly audiences. Approximately 45 hours (500,000 words) of the corpus are fully annotated. As discussed above, the morphological and phonological coding used for R-JMICC is based on CSJ. Thus the two corpora use the same coding criteria. For our analysis, we selected the simulated speech data of female speakers who are in a similar age range as the R-JMICC (in their 30s). There were 10 participants that satisfied the criteria. Three of the participants contributed three different recording samples, making the total recording samples 16. Each sample consisted of approximately 10 minutes of simulated speech, 1,300 LUW on average.

2.2.3 Results from Spoken Corpora

In a survey, where participants are asked to list words, mostly content words in citation forms are reported. In a corpus, in contrast, function words and inflected forms are mixed. To deal with this, we analyzed the corpus in two ways. First, we analyzed all Long Unit Words (LUW) in the corpus. We will call this the full corpus analysis. LUW counts function words as a unit, i.e., case markers, particles and auxiliaries are counted as one LUW. The infant-directed speech of R-JMICC should be a reasonable sample of the language input infants receive in real life. Second, we selected nouns (excluding proper nouns and numerals) from the corpus and analyzed them separately. This should be more comparable to the IDV survey than analyzing every word in the corpus, although verbs and adjectives are excluded from the analysis because of inflection.

For each speaker, the frequency of LUW was calculated by length in morae. Each LUW was also classified by its syllable weight depending on the position of special morae in the word. For each speaker, we calculated the proportion of RSR, RRR and other word types among 3-mora words.[2] Similarly, the proportion of RSRS, RRRR, and other types of words among 4-mora words were calculated for each speaker.

Figure 2 shows the distribution of items in terms of mora length from all the Japanese data we analyzed in this paper. When all words in the corpora were analyzed, 1- and 2-mora LUW occurred much more frequently than other types of words. One and two-mora LUW accounted for 34% and 33% of R-JMICC IDS, 37% and 35% of ADS, and 39% and 27% of CSJ. This was because many function morphemes, such as case particles, post-positions and auxiliaries tend to be 1 and 2 mora long. When nouns were selected for analysis, much fewer one-mora words occured (3.7% of IDS, 5.6% of ADS, 1.6 % of CSJ). Yet, the proportion of 2-mora words is higher (15% of IDS, 28% of ADS, 29% of CSJ) than other samples, suggesting that the spoken corpora tend to include shorter words than other samples. Interestingly, adult speech samples contained higher rates of 2-mora words than ID speech (one-way within-subject ANOVA between AD vs ID in R-JMICC corpus, $F_{1,21} = 152.00$, $P < 0.0001$) suggesting that the frequent use of 2-mora words in adult speech is not related to talking to infants. In both all-word analysis and noun-only analysis of the three samples, approximately 90% of 2-mora words contain no special mora, i.e., they were bisyllabic, light-light syllable words, while 10% of them contained a special mora, i.e., monosyllabic words with a heavy syllable.

Figure 3a shows the proportion of 3-mora words that were RRR, RSR, or other types, and Figure 3b shows 4-mora words that were either RRRR, RSRS or other types. In the full corpus analysis, the occurrence of RSR type words among 3-mora words, and the occurrence of RSRS type words among 4-mora

[2] In the full corpora, inflected forms of LUW could take forms that are not possible in words in isolation, e.g., /Qpanashi/ (SRRR). For the purpose of this paper, they are included in "others."

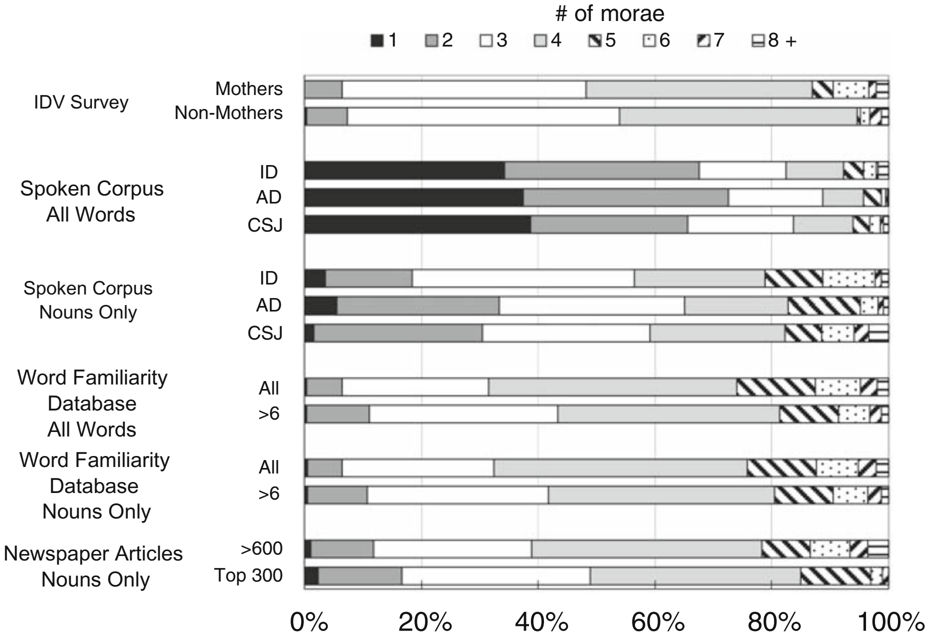


FIG. 2. Distribution of word length in terms of number of morae.

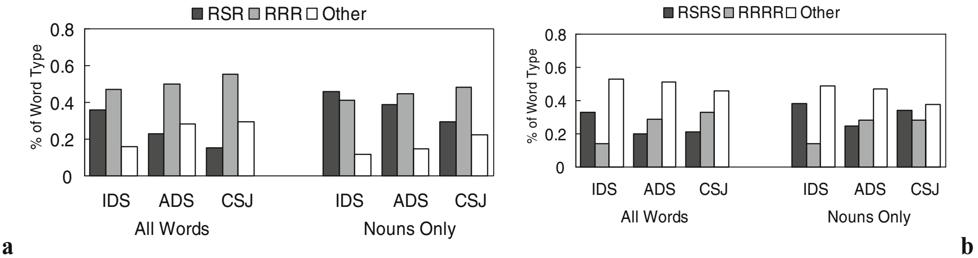


FIG. 3. Spoken Corpora: word types (**a**) among 3-mora words and (**b**) among 4-mora words.

words were significantly higher in ID speech than AD speech ($F_{1,21} = 31.74$, $P < 0.0001$; $F_{1,21} = 19.44$, $P < 0.0002$, respectively). When only nouns are selected for analysis, the occurrence of RSR type words was only marginally more frequent in ID speech than AD speech ($F_{1,21} = 3.08$, $P = .09$) while RSRS type words were significantly more frequent in ID speech than AD speech ($F_{1,21} = 8.86$, $P < 0.01$).

Among 3-mora words, RRR was the most frequent word type in all three samples when full corpora were analyzed. The ratio of RSR words compared to the other two types was highest in IDS, followed by ADS, and CSJ. When only nouns were analyzed, RSR type words occurred more frequently than RRR type words in IDS, while RRR type words occurred more frequently in ADS and CSJ. In full corpora analysis 4-mora words, other types (this category includes 3-syllable words, RSRR, RRSR, RRRS, 2-syllable words RSSR and RRSS, and other forms) occurred more frequently in all three corpora. In IDS, RSRS words occurred more frequently than RRRR words, while the opposite pattern was observed in ADS and CSJ. In noun-only analysis, RSRS type words occurred most frequently in the CSJ corpus. In IDS and CSJ, RSRS type words occurred more frequently than RRRR type words, while in ADS speech, RSRS and RRRR type words occurred at an approximately equal frequency.

Taken together, the prosodic forms dominantly reported in the IDV survey, i.e., RSR and RSRS type words are found to occur more frequently in infant-directed speech than in adult-directed speech. But it is not the case that these two forms dominate infants' input overwhelmingly. In addition, the comparison between ID speech and AD speech shows different patterns for RSR words and RSRS words. The frequent occurrence of RSR words seems to be specific to IDS as it does not occur as dominantly in the two adult speech corpora. Interestingly however, RSRS type words occurred frequently in CSJ nouns.

2.3 NTT Word Familiarity Database

To further investigate the difference and similarities between infant-directed speech and adult Japanese language, we analyzed two additional data sets. First, we asked whether or not the IDV forms are representative of Japanese vocabulary that is highly familiar to adult Japanese speakers. It could be that the IDV forms are not unique to infant-directed speech, but they are representative of adult Japanese vocabulary anyway, and the frequent occurrence of them in infant-directed speech is an exaggeration of this tendency. For this analysis, we used the NTT Database Series, Lexical Properties of Japanese: Word Familiarity (henceforth referred to "Word Familiarity Database", Amano and Kondo 1999), which compiled Japanese adults' ratings of how familiar a word is for all content words in a published dictionary. Second, we analyzed the word frequency compiled from newspaper articles, as described in section 2.4 below.

2.3.1 Method

The Word Familiarity Database (Amano and Kondo 1999) provides familiarity ratings by adult Japanese native speakers on all of the content words that are listed in a published Japanese dictionary (Shinmeikai Dictionary, total of 69,084 content words). We selected Japanese content words that are rated as highly familiar by Japanese adults, (rated higher than 6 on a 7-point-scale, in simultaneous visual/auditory presentation condition, total of 4,305 words). To be compatible with the analysis of nouns in the corpora and the newspaper articles, we also selected nouns only for a separate analysis (total number of nouns in the database; 26,019, nouns rated higher than 6; 1,378 words).

As common practice for a dictionary, the only phonological coding available for the items in the Word Familiarity Database is "kana" transcription, which is mora-based. Syllable weight has to be derived from the position of special morae. For long vowels, the database provide the ratings by native speakers as to whether or not they think it is pronounced as a long vowel for combinations of vowels (e.g., /ei/, /ou/ etc). We counted a vowel combination as a long vowel when it was judged as a long vowel by more than 50% of the raters. However, there is no coding for whether or not a vowel sequence is pronounced as a diphthong, and diphthongs are not counted as special morae in our analysis of the database. This resulted in underestimating the number of special morae and overestimating the number of regular morae.

2.3.2 Results of Word Familiarity Database Data

Figure 2 indicates the distribution of content words in terms of mora length from the Word Familiarity Database. It shows that 4-mora words are the most frequent, followed by 3-mora words both when we include all words from Shinmeikai dictionary, and also when only those rated higher than 6 on the familiarity rating are included. The result was very similar when only nouns were analyzed. Figure 4a shows the proportion of 3-mora words according to their syllable weight, and Figure 4b shows that of 4-mora words. Among 3-mora words, the RRR type was by far the most frequent type, while RSR words and other types occurred much less frequently. RRR type words occurred less frequently among nouns. But they were still much more frequent than RSR type words. Among 4-mora words, the RSRS type did not occur as frequently as the other two types when all words were analyzed. RSRS words occurred more often than RRRR type when only nouns were analyzed. This tendency became more pronounced among highly familiar nouns.

These results from the Word Familiarity Database show a different pattern from the IDV survey. Among three-mora words, RRR words occurred much more frequently than RSR words in all of the analysis—either when we analyzed all words or among highly familiar words, all nouns or highly familiar nouns. Thus, it seems reasonable to assume that the RSR form is not a typical 3-mora word in adult Japanese content words. Among 4-mora words, slightly different

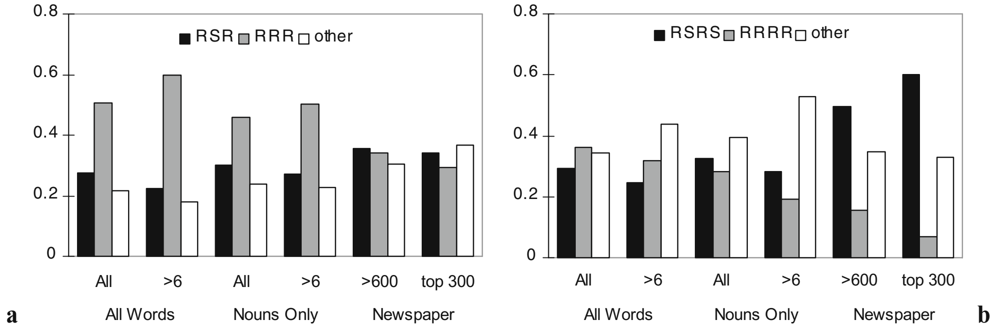


FIG. 4. Word Familiarity Database: proportion of (**a**) 3-mora words and (**b**) 4-mora words according to syllable structure.

patterns emerged when we analyzed all words or just nouns. When all content words were included in the analysis, RSRS type words occurred no more frequently than RRRR words. Among nouns, in contrast, the proportion of RSRS type words was higher than RRRR type words, and this tendency was stronger among highly familiar 4-mora nouns. As discussed later, the same pattern was found among 4-mora nouns in CSJ corpus and newspaper articles. In sum, it is unlikely that the frequent use of IDV forms, at least for RSR words, is an "exaggeration" of existing adult Japanese vocabulary.

2.4 Newspaper Articles

In addition, we also analyzed the word frequency data from a newspaper corpus. So far, we have analyzed three types of Japanese adult data sets; two spoken corpora, (R-JMICC AD speech and CSJ) and the Word Familiarity Database. Newspaper articles supplement the other data as it represents formal writing, and might show a different pattern of dominant prosodic forms. For this analysis, we used the NTT Database Series, Lexical Properties of Japanese: Word Frequency (Amano and Kondo 2000). This database compiles 14 years of Asahi Newspaper articles, between 1985 and 1998. The total size of the corpus is approximately 300 million words in text, and the total number of wordtypes contained in the article was 341,771. From this, we selected common nouns that appeared at least 600 times in the text (total of 12,292 words). We did not include verbs or adjectives in our analysis for the same reason as the spoken corpora. Proper nouns, numerals and the category of "Keishiki Meishi[3]" are also excluded from our analysis.

As in other data, words were tabulated according to their length in mora, and 3- and 4-mora words were cross-tabulated in terms of their syllable weights. For this purpose, we hand-coded 3- and 4-mora nouns for long vowels and diphthongs. The second vowels of two-vowel sequences are coded as long vowels or diphthongs when a native speaker coder judged that they would be typically pronounced as long vowels or diphthongs in a given word. The bottom two bars of Figure 2 show the data from the newspaper articles. As can be seen in this figure, 3- and 4-mora words were the most frequent words. Figure 4a and 4b show the distribution of 3- and 4-mora words in terms of their syllable weights. Among 3-mora words, RSR, RRR and other types of words occurred at an approximately equal frequency. Among 4-mora words, however, RSRS type words accounted for almost half of the items, and for 60% of the top 300 items.

The results suggest that in newspaper articles, RSR words among 3-mora words are not particularly dominant, while RSRS type words occur frequently among nouns. In the CSJ corpus and Word Familiarity Database discussed above, RSRS words were also found frequently among 4-mora nouns. RSRS might be a dominant syllable structure for 4-mora nouns in adult Japanese.

[3] In this database, various types words such as "atari" (near), "totan" (as soon as), "ori" (when) are classified into this category. We excluded them from our analysis since it is not clear what parts of speech they ought to be classified.

3 General Discussion

In this paper, we investigated the prosodic form of infant-directed vocabulary reported by Japanese mothers. We found that two forms occur highly frequently in IDV; RSR (HL) words among 3-mora words and RSRS (HH) words among 4-mora words. In comparing the IDV data to other Japanese data, we reported three main results: 1) The IDV survey revealed remarkable similarities between the results of mothers and non-mothers in every measure we compared. 2) The two dominant IDV forms occur significantly more frequently in infant-directed speech than in adult-directed speech, but their occurrence in actual speech is not as dominant as in the IDV survey. 3) The frequent occurrence of IDV forms is not a general property of adult Japanese speech. In particular, the RSR (HL) form among 3-mora words was never found to occur dominantly in any of the adult Japanese data we analyzed. RSRS words, in contrast, were found to occur dominantly among 4-mora nouns in CSJ, Word Familiarity Database and Asahi newspaper articles, but not when all words were analyzed.

3.1 Where Do the IDV Forms Come from?

A critical issue we need to address is where these forms come from. We will restrict our discussion to RSR (HL) words here. The question arises from the fact that most of the IDV forms are, at least on the surface, phonologically unrelated to the adult form—/buHbu/ (car) is not phonologically related to the adult form /kuruma/. Japanese infants, who learn these forms initially, would have to unlearn it and relearn the adult form later. The presence of word learning constraints, such as “mutual exclusivity” (c.f., Markman 1994), could mean that the use of IDV may hinder Japanese children’s word learning. In other words, the RSR words are doubly dissociated from the target language—phonologically dissociated from the target word at the level of individual words, and distributionally dissociated from adult Japanese vocabulary. Then, why do Japanese mothers use such forms? In the literature, several explanations have been put forward. We will consider three of them below.

3.1.1 Exaggerated Form of Adult Japanese

It is possible that the prosodic form of the IDV words are reflecting general properties of the Japanese language, even though the individual IDV items may be phonologically unrelated to their target words. To address this issue, we analyzed adult Japanese in a number of formats; two kinds of spoken corpora, a Word Familiarity Database, and a newspaper corpus. In the analysis of the Word Familiarity Database, RSR forms were not found to occur frequently among highly familiar words. As we discussed above, our analysis of the Word Familiarity database underestimates the occurrence of special morae, as diphthongs were not counted as special morae. However, even taking into account the underestimation of diphthongs, the results showed that RSR forms are not cited very frequently among “highly familiar” words.

In terms of frequency of occurrence in written and spoken corpora, we saw that the same pattern was found. RRR words occurred more frequently than RSR words in R-JMICC AD corpus and CSJ, and RSR forms occurred approximately as often as RRR or other forms in the Asahi newspaper corpus. In no case did we find RSR types more dominant than RRR or other types. Taken together, we have no data that suggest that the frequent occurrence of RSR type words among 3-mora words is the general pattern found in adult Japanese.

3.1.2 Default Form for New Word Formation

Itô and Mester (1992) discussed that when new words are created by shortening, or combining loan words, the resulting form often takes the form of RSR (HL). For example, when "paNhureQto" (pamphlet) is shortened, it becomes "paN. hu". Similarly, in a word game where the first half and second half of a word are inverted, the resulting word often takes the RSR or RSRS form. For example, "jazu" (jazz) will become "zuHja" (RSR) instead of simply juxtaposing the two syllables, i.e., "zuja". Kubozono (1995; 2002; 2006) argues that the RSR form in IDV may be derived by the same principle. It is possible that for Japanese mothers, producing IDV is like creating new words to fit the need of communicating with their children. Consequently, IDV words produced by mothers would share the same forms as the new word formation.

Data from the present paper are consistent with this interpretation. That is, the default word forms for creating new words and IDV forms share the same prosodic forms. It does not mean, however, that this hypothesis implies a causal relation between the two—just because the typical IDV forms share the same prosodic forms with templates for creating new words does not mean one is the cause of the other. It is very possible that the two types of word forms are constrained by the same principle. But we are still left with explaining why these specific forms are preferred for IDV (and for new word formation), because they are not the general pattern in adult Japanese vocabulary.

3.1.3 Imitation of Early Production by Children

Perhaps the most frequent explanation that is offered for the specific form of IDV is that it is attributable to children's early production. For example, Murata (1960; 1968) argued that the specific forms of the infant-directed vocabulary come from children's early production. That is, these are the forms children produce at an early stage of word production, and the mothers are adopting them in their own production. Murase et al. (1992), and Ogura et al. (1993) found that the use of the specialized form vocabulary decreases as children's production of adult-form vocabulary increases, suggesting that there is an interaction between the mother's use of these forms and children's vocabulary development.

[4] We thank Mitsuhiko Ota for allowing us to access his original coding of the three children's utterances.

In order to evaluate this explanation, we analyzed the production data from Ota (2006) in terms of their syllable weights.[4] The original data was taken from the Miyata corpus (Miyata 1992; 1995; 2000) of Japanese CHILDES (Miyata 2004). We used the three children's natural utterances at age 2;1. Figure 5 shows the proportion of children's utterances in terms of mora length. Calculations are based on token count for each child. At this age, children's utterances were already mostly 2, 3 or 4 mora long, which is comparable to the word length in adult spoken corpora shown in Figure 2. We then classified 3 and 4-mora words according to their syllable weights. As shown in Figure 6, the pattern of their production is very similar to the IDV survey results, that is among 3-mora words, a very high proportion of them were of RSR (HL) form for all three children, and for 4-mora words also, RSRS (HH) forms were produced much more frequently than RRRR type words for all three children. Children produced higher proportion of "other" types of 4-mora words than the IDV survey. Thus, children's early productions indeed show the predominant frequency of RSR and RSRS patterns.

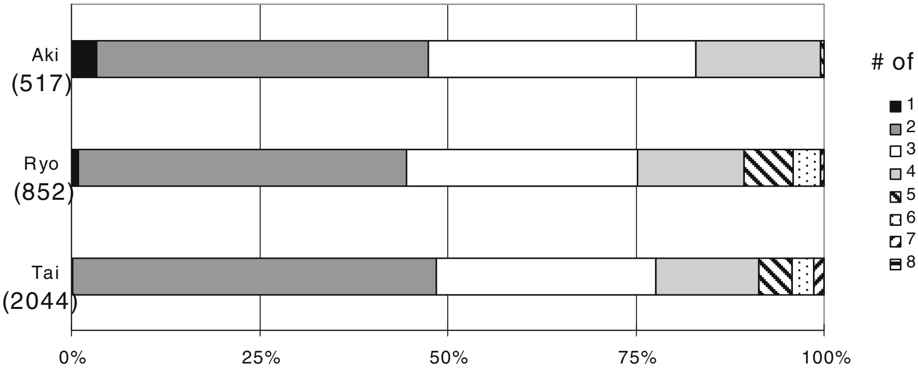


FIG. 5. Children's production at 2;1 (Reanalysis of data from Ota, 2006.) Proportion of words in terms of mora-length (# of tokens analyzed).

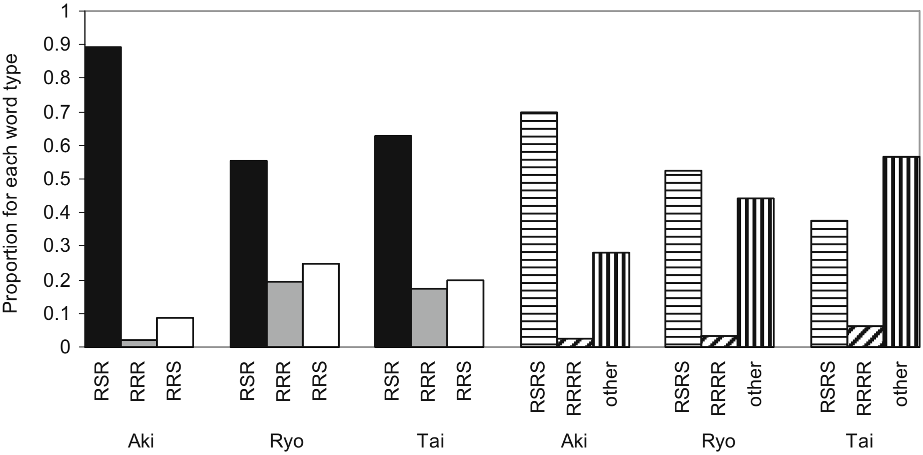


FIG. 6. Children's productions of 3 and 4-mora words.

At the same time, the fact that the IDV forms are similar to the early form of children's production does not necessarily support a claim that mothers (and other adults) are "adjusting" their speech imitating children's production. If so, we predict that mothers who had more experience with children at an early production stage should be using more of these forms than non-mothers or mothers of younger infants who are not producing words yet. As shown above, however, our data showed that non-mothers with little experience with children produced a remarkably similar list of IDV as mothers did. In addition, the mothers in our survey have not had much experience listening to their own child's production. Except for four mothers who also had older children, mothers in our survey had 8 to 12 month old infants as their only child. Since infants at this age do not have much productive vocabulary, it is not likely that the mothers' list in the survey was inspired by their own children's production. Our data does not support a direct version of learning-hypothesis where individual mothers are assumed to learn to produce the RSR forms of IDV by listening to the actual utterances of the children. It does not rule out the possibility that more indirect forms of learning can occur through media or other cultural experiences about how young children's utterances would sound like, and it could influence the IDV form indirectly.

3.2 Four-Mora Nouns

We need to examine the RSRS forms among 4-mora words more carefully. The RSRS type words were highly dominant in IDV survey, and we found that it also occurred in ID speech of R-JMICC. But it was not found to occur dominantly in AD speech when we analyzed all words in the corpus or selected just nouns. In contrast, when we selected nouns in CSJ, Word Familiarity Database, and the newspaper articles, we found that RSRS type words occur more dominantly than other types. The pattern was stronger in CSJ and newspaper articles, compared to the Word Familiarity Database. This tendency did not seem to change, whether we analyzed highly familiar nouns or all nouns in the Word Familiarity Database, and highly frequent nouns or all nouns in the newspaper articles. The fact that this pattern was found from the samples derived from broader range of adult vocabulary, but not when talking to an experimenter about limited topics on child rearing, suggests that this may be a general pattern of 4-mora Japanese nouns in relatively sophisticated context.

As discussed in section 2.1.2.2, many of the RSRS words in IDV are derived from onomatopoeic or mimetic expressions, involving a duplication of heavy syllables, or two light syllable units. RSRS type words in adult nouns, on the other hand, are mostly derived from two-Chinese character compounds, where the Chinese pronunciation of each character corresponds to a heavy syllable, e.g., "giNkoH" (bank), "kaQtoH" (conflict). Thus, even though their prosodic forms may be similar, RSRS words among adult nouns are quite distinct lexical items from IDV RSRS words. Nonetheless, it is interesting that two sets of vocabulary that arise from very different sources share the same prosodic form, i.e., Heavy-Heavy bisyllabic form.

3.3 Conclusion

In this paper, we showed that Japanese mothers use a specialized vocabulary when talking to children, and that there are two prosodic forms that are predominant—the 3-mora, RSR, Heavy-Light (HL) type, and the 4-mora, RSRS, Heavy-Heavy (HH) type. We evaluated three hypotheses that attempt to explain why these forms are used—1) an exaggeration of a pattern also found in adult Japanese, 2) the default form for new word formation, 3) adaptation to children's early production. We did not find support for the first hypothesis, but we found that the default form for new word formation and children's early production both share the same patterns. As discussed above, neither of these forms may be causally related to IDV forms. But Japanese native speakers seem to know what prosodic forms would make good Japanese infant-directed vocabulary as a part of their phonological knowledge. Given the evidence that these are not the general property of adult Japanese, we are still left to account for why these forms are favored in IDV (new word formation, and in early production).

References

Abercrombie D (1967) Elements of general phonetics. Edinburgh University Press, Edinburgh

Amano S, Kondo T (1999) Nihongo-no Goi-Tokusei (Lexical properties of Japanese) vol 1. Sanseido Shuppan, Tokyo

Amano S, Kondo T (2000) Nihongo-no Goi-Tokusei (Lexical properties of Japanese) vol 7. Sanseido Shuppan, Tokyo

Cooper RP, Aslin RN (1990) Preference for infant-directed speech in the first month after birth. Child Development 61:1584–1595

Cutler A, Carter DM (1987) The predominance of strong initial syllables in the English vocabulary. Computer Speech and Language 2:133–142

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

Fernald A, Kuhl P (1987) Acoustic determinants of infant preference for motherese speech. Infant Behavior and Development 10:279–293

Fernald A, Mazzie C (1991) Prosody and focus in speech to infants and adults. Developmental Psychology 27:209–221

Fernald A, Morikawa H (1993) Common themes and cultural variations in Japanese and American mothers' speech to infants. Child Development 64:637–656

Fernald A, Simon T (1984) Expanded intonation contours in mothers' speech to newborns. Developmental Psychology 20–1:104–113

Fernald A, Taeschner T, Dunn J, Papoušek M, de Boysson-Bardies B, Fukui I (1989) A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language 16:477–501

Fisher C, Tokura H (1996) Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. Child Development 67:3192–3218

Grieser D, Kuhl P (1988) Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. Developmental Psychology 24:14–20

Hayashi A, Tameka Y, Kiritani S (2001) Developmental change in auditory preferences for speech stimuli in Japanese infants. Journal of Speech, Language, and Hearing Research 44:1189–1200

Igarashi Y, Mazuka R (2006) Hahaoya tokuyuu no hanashikata (mazariizu) wa otona no nihongo to doo chigauka (How do "motherese" differ from adult Japanese?). The Institute of Electronics, Information and Communication Engineers, Technical Report NLC2006-34,SP2006-90(2006-12):31–35

Itô J, Mester A (1992) Weak linking and word binarity, *Linguistic Research Center, LRC-92-09, University of California, Santa Cruz*. In: Honma T, Otazaki M, Tabata T, Tanaka S (eds) A new century of phonology and phonological theory. A festschrift for professor Shosuke Haraguchi on the occasion of his sixtieth birthday. Kaitakkusha, Tokyo, pp 26–65

Kelly M, Martin S (1994) Domain general abilities applied to domain specific tasks: Sensitivity to probabilities in perception, cognition, and language. Lingua 92:105–140

Kubozono H (1995) Gokeisei to Oninkoozoo (Word formation and prosodic structure). Kuroshio Shuppan, Tokyo

Kubozono H (1999) Mora and syllable. In: Tsujimura N (ed) The handbook of Japanese linguistics. Blackwell, Oxford, pp 31–61

Kubozono H (2002) Prosodic structure of loanwords in Japanese: Syllable srucutre, accent and morphology. Journal of the Phonetic Society of Japan 6-1:79–97

Kubozono H (2006) Yoojigo no oninkoozoo (Prosodic structure of children's words). "*Gengo*", Taishuukan 35-9:28–35

Kuhl P, Andruski J, Chistovich I, Chistovich L, Kozhevnikova E, Ryskina V, Stolyarova E, Sundberg U, Lacerda F (1997) Cross-language analysis of phonetic units in language addressed to infants. Science 277:684–686

Ladefoged P (1975) A course in phonetics. Harcourt Brace Jovanovich, New York

Maekawa K (2006) Gaisetsu (General introduction). In: Construction of the corpus of spontaneous Japanese. National Institute for Japanese Language, Report 124

Markman E (1994) Constraints on word meaning in early language acquisition. Lingua 92:199–227

Mazuka R, Igarashi Y, Nishikawa K (2006) Input for learning Japanese: RIKEN Japanese Mother-Infant Conversation Corpus. The Institute of Electronics, Information and Communication Engineers, Technical Report TL2006-16 (2006-07):11–15

Miyata S (1992) Wh-Questions of the third kind: The strange use of wa-questions in Japanese children. Bulletin of Aichi Shukutoku Junior College 31:151–155

Miyata S (1995) The Aki corpus. Longitudinal speech data of a Japanese boy aged 1.6–2.12. Bulletin of Aichi Shukutoku Junior College 34:183–191

Miyata S (2000) The Tai corpus: Longitudinal speech data of a Japanese boy aged 1;5.20–3;1.1. Bulletin of Aichi Shukutoku Junior College 39:77–85

Miyata S (2004) Kyookara Tsukaeru Hatsuwa Deetabeesu CHILDES Nyuumon (An introductory manual for to CHILDES database). Hitsuji shoboo, Tokyo

Murase T, Ogura T, Yamashita Y (1992) Ikujigo no kenkyuu (1). Doobutsu meishoo ni kansuru hahaoya no shiyoogo: Ko no geturei ni yoru chigai (A study of child-rearing vocabulary (1). Mothers' use of animal terms: Effects of children's age). Annual Bulletin of Shimane University, Faculty of Education 17:37–54

Murata K (1960) Ikugijo no kenkyuu (A study of child-rearing vocabulary). Shinrigaku Kenkyuu 31:33–38

Murata K (1968) Youji no Gengo Hattatsu (Children's language development). Baihuukan, Tokyo

Newport E, Gleitman L, Gleitman H (1977) Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In: Snow C, Ferguson C (eds) Talking to children, language input & acquisition. Cambridge University Press, Cambridge, pp 109–149

Ogura T, Yamashita Y, Murase T, Dale PS (1993) Some findings from the Japanese early communicative development inventory. Annual Bulleting of Shimane University, Faculty of Education 27:27–39

Ota M (2006) Input frequency and word truncation in child Japanese: Structural and lexical effects. Language and Speech 49:261–295

Otake T (2006) Speech segmentation by Japanese listeners: Its language-specificity and language-universality. In: Nakayama M, Mazuka R, Shirai Y (eds) Handbook of East Asian psycholinguistics, vol II, Japanese. Cambridge University Press, Cambridge, pp 201–207

Otake T, Hatano G, Cutler A, Mehler J (1993) Mora or syllable? Speech segmentation in Japanese. Journal of Memory and Language 32:258–278

Otake T, Hatano G, Yoneyama K (1996) Japanese speech segmentation by Japanese listeners. In: Otake T, Cutler A (eds) Phonologcal structure and language processing: Cross-linguistic studies. Morton de Gruyter, Berlin, pp 183–201

Port R, Dalby J, O'Dell M (1987) Evidence for mora timing in Japanese. Journal of the Acoustical Society of America 81-5:1574–1585

Snow C, Ferguson C (1977) Talking to children: Language input & acquisition. Cambridge University Press, London

Tomonaga K (1997) Zenkoku Yoojigo Jiten (National child-vocabulary dictionary). Tokyodoo Shuppan, Tokyo

Trainer L, Austin C, Desjardins R (2000) Is infant-directed speech prosody a result of the vocal expression of emotion? Psychological Science 11:188–195

Warner N, Arai T (2001) Japanese mora-timing: A review. Phonetica 58:1–25

5
Short-Term Acoustic Modifications During Dynamic Vocal Interactions in Nonhuman Primates—Implications for Origins of *Motherese*

Hiroki Koda

1 Dynamic Vocal Interactions in Primate Species

1.1 Vocal Exchanges in Animals and Humans

In ethology, communication, including human language or conversation, is defined as the interaction between two individuals, a *sender* and a *receiver*, using a *signal*. The sender produces a signal that conveys information. The signal and information are transmitted through the environment and are detected by the receiver, who uses the information to help in deciding how to respond. The receiver's response affects the fitness of both the sender and the receiver (Bradbury and Vehrencamp 1998). Information exchange is an essential component of the definition of animal communication. In *true communication*, both sender and receiver benefit from the information exchange (e.g., Marler 1977).

Vocal exchange is a characteristic communication style in which a sender produces a vocalization to address a receiver, and the receiver emits a call in response within a short-term interval. Many studies of vocal communication have emphasized vocal exchanges in birds (Beecher et al. 1996; Krebs et al. 1981; McGregor et al. 1992), group-living mammals (Janik 2000; Miller et al. 2004), and nonhuman primates (Biben et al. 1986; Koda 2004; Masataka and Biben 1987; Snowdon and Cleveland 1984; Sugiura 1993; 1998; 2001; 2007). Human conversation, which is organized on the basis of taking turns in the role of speaker, is also included in this communication style (e.g., Goodwin 1981). Primatologists and physical anthropologists have paid special attention to the issue of language evolution, and comparative analyses between nonhuman and human vocal communications have been attempted from ethological, psychological, and ecological views (for review, see Ghazanfar 2003). Vocal exchanges in some nonhuman primate species are achieved by taking turns in the roles of sender and receiver with precise

Primate Research Institute, Kyoto University, Kanrin 41, Inuyama, Aichi 484-8506, Japan

temporal regularity; therefore, the pattern of this type of vocal exchange communication is somewhat conversational and analogous to human conversation.

A fairly strict alternation of calling can take place between two individuals, with the two participants exhibiting extraordinary precision in the call exchange vocalizations. Antiphonal calling behavior with precise temporal regularity has been reported in several primate species (squirrel monkeys, Masataka and Biben 1987; Japanese macaques, Sugiura 1993; 2007; siamangs, Geissmann 1999). In particular, conversational analysis has been conducted intensively for the vocal interactions of wild Japanese macaques. Regarding the temporal structure of call exchanges in wild Japanese macaques, Sugiura (1993) found that macaques vocally interact and exchange calls with other group members with short inter-call intervals. Moreover, Sugiura (2007) recently confirmed that macaques can flexibly adjust the timing of vocalizations in accordance with preceding utterances from other group members. Temporal regularity is a fundamental ability required for vocal exchange communication.

1.2 Acoustic Modification During Vocal Exchange in Nonhuman Primates

Although previous studies of vocal exchanges in both wild and captive monkeys have shown a high ability for temporal adjustment of the vocal production (e.g., Gilissen 2004), there is little evidence that monkeys possess vocal plasticity, i.e., the ability to modify the acoustic features of the vocalization. Early studies of deafening, social isolation, and cross-fostering experiments in immature monkeys did not show vocal learning of their acoustic features, resulting in a lack of vocal plasticity over a long time scale (Hammerschmidt et al. 2000; 2001; Owren et al. 1992; Winter 1969). Moreover, evidence of the neural structures underlying vocal control suggests that monkeys cannot voluntarily modify the acoustic structure of the call independent of their emotional status, which also means a lack of vocal plasticity over a short time scale (e.g., Deacon 1989). Vocal plasticity is one of the most remarkable attributes separating vocal communication between human and nonhuman primates (for review, see Egnor and Hauser 2004).

However, recent studies have indicated a greater plasticity of vocal production in nonhuman primates than previously considered. First, evidence has accumulated of acoustic variation between wild social groups, termed “vocal dialects” (Barbary macaques, Fischer et al. 1998; rhesus macaques, Hauser 1991; 1992; chimpanzees, Crockford et al. 2004; Marshall et al. 1999; Mitani et al. 1992; Japanese macaques, Sugiura et al. 2006; Tanaka et al. 2006). Particularly, recent studies of wild Japanese macaques have strongly suggested that the physical environmental properties of the habitat acoustically influence population differences in the call structures (Sugiura et al. 2006; Tanaka et al. 2006). These findings imply some extent of vocal plasticity over a relatively long time scale. Second, the phenomenon of immediate acoustic modification of primate calls has been reported in captive subjects (Brumm et al. 2004; Egnor and Hauser 2006;

Egnor et al. 2006; Masataka and Symmes 1986; Schrader and Todt 1993). Furthermore, some reports have suggested immediate acoustic modification in vocal exchange communication in wild Japanese macaques (Koda 2004; Sugiura 1998) and chimpanzees (Mitani and Brandt 1994). Generally, the recent findings in both captive and wild subjects also showed vocal plasticity over a relatively short time scale. Although most previous studies investigating the vocal plasticity of nonhuman primates do not deny the limitations of primate vocal production, the contrary evidence challenges the traditional view of the restricted modifiability of vocalizations in nonhuman primates.

In this chapter, I review the phenomenon of short-term acoustic modification of the vocalizations of wild Japanese macaques. Here, I report the accumulating evidence that wild Japanese macaques immediately modify the acoustic features of their contact calls according to the preceding context in which a macaque has emitted a contact call and fails to elicit a response call from another group member. Moreover, in an attempt to examine what properties are shared in vocal exchange communication between human and nonhuman primates, I compared patterns of human mother–infant vocal interactions to vocal interactions between Japanese macaques.

2 Short-Term Acoustic Modifications in Wild Japanese Macaques

2.1 Coo Calls of Wild Japanese Macaques

The contact call is one of the call types used by social-living animals. Primates usually organize a wide variety of social groups and coordinate their behaviors within a group (Smuts et al. 1987). Vocal exchanges of contact calls serve to coordinate with other group members. Contact calls are produced in various calm contexts, e.g., while feeding and moving, and are the most frequent intragroup vocalizations; they have been researched in several primate species (e.g., Japanese macaques, Itani 1963; pygmy marmosets, Cleveland and Snowdon 1982; squirrel monkeys, Boinski and Mitchell 1992; chacma baboons, Cheney et al. 1996; Rendall et al. 2000; rhesus macaques, Rendall et al. 1996).

Studies of vocal communications in Japanese macaques, *Macaca fuscata*, have also concentrated on the contact call or vocal exchange communication. Previous studies of the call frequently emitted by adult and juvenile Japanese macaques during calm states have referred to it as the "coo call" (e.g., Green 1975; Fig. 1). In particular, the conversational analysis of coo calls on Yakushima Island, Japan, has shown some striking evidence of vocal flexibility. Sugiura (1993) observed coo call exchanges and revealed a communication rule with precise temporal regularity. He also reported the vocal matching phenomenon in which the responding caller matches the acoustic feature of the response coo call to that of the initial call of the preceding caller during the vocal exchange (Sugiura 1998). Moreover, Tanaka et al. (2006) revealed acoustic population differences through both longitudinal and cross-sectional developmental research. These differences are likely to be

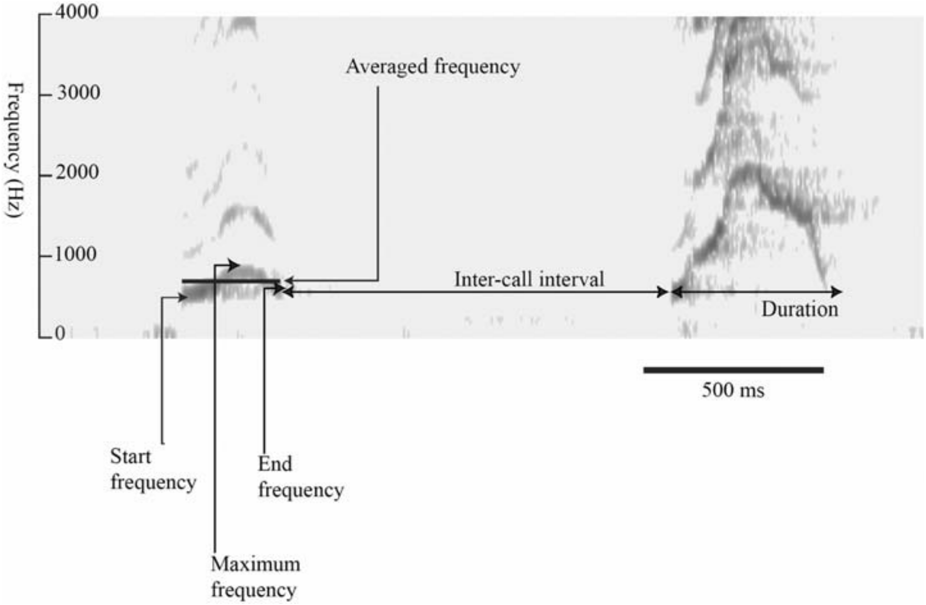


FIG. 1. Representative sound spectrograph of two successive coo calls emitted by the same caller indicating the acoustic properties measured for the analysis. In a previous report, these sequences were defined as self-repeated sequences (modified from Koda 2004, with permission).

caused by environmental factors (Sugiura et al. 2006). Consequently, wild Japanese macaques show acoustic flexibility over both short and long timescales in their dynamic communication. Such vocal flexibility of the contact call is likely to contribute to the maintenance of efficient vocal exchanges.

2.2 Repeated Calling Attempts by Wild Japanese Macaques

The conversational temporal rules of contact call exchanges have been shown in several primate species (Koda 2004; Masataka and Biben 1987; Sugiura 1993; 2007). In particular, the phenomenon of repeated utterances of the contact call was described. Previous studies of the conversational analysis of vocal exchanges in wild Japanese macaques have reported that when there is no response to the contact call of an individual within a short interval, the animal then repeatedly vocalizes contact calls (Koda 2004; Sugiura 1993; 2007). These repeated vocalizations are considered effective in attracting the attention of other group members. In the previous studies, the timing of consecutive calls was conversationally analyzed to define the patterns of repeated calling attempts (Koda 2004; Sugiura 1993; 2007).

In the conversational analysis of coo call exchanges, the inter-call intervals between two consecutive coo calls were measured, and the overall distribution

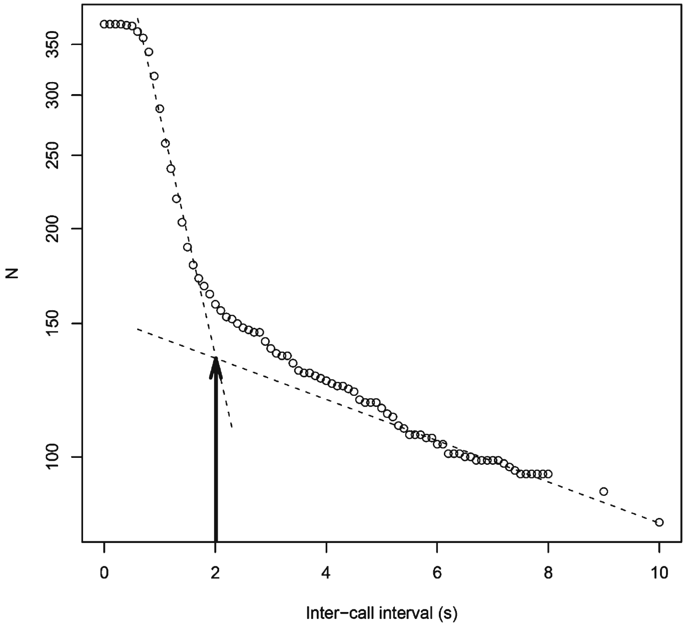


FIG. 2. Log-survival plot of the inter-call intervals between two consecutive coo calls emitted by the same subject. The arrow indicates the abrupt change in curvature and the criterion interval chosen for the repeated addressing attempt (modified from Koda 2004, with permission).

of the inter-call intervals was generated as a log-survival plot. Figure 2 shows the overall distribution of the timing of consecutive coo calls observed in a previous study (Koda 2004). If calls are emitted randomly, the distribution of time intervals is expected to be exponential, and the graph should form a straight line with a negative slope (Sibly et al. 1990). However, the curve does not show a steady decrease, but instead abruptly changes in slope angle at 2 s. This change indicates that the overall distribution of the call sequence has two distinctive patterns. One major pattern involves the focal monkey repeating a coo call within a short interval of the preceding call, whereas the other minor pattern involves the monkey repeating a call after a relatively longer interval. In the former pattern, the subsequent call seems to be related to the preceding call. Two successive coo calls in which the inter-call interval was less than 2.0 s were interpreted as a repeated attempt of coo calls addressing other group members. In a study of inter-call intervals, Koda (2004) proposed that these repeated attempts be defined as *self-repeated sequences* (SRSs) of the two consecutive coo calls, and the first and second SRS calls were defined as the initial and repeated coo calls, respectively. Figure 1 illustrates a SRS.

2.3 Pattern of Self-Repeated Sequences during Vocal Exchanges

An acoustic analysis of the initial and repeated coo calls in the SRS indicated characteristic acoustic modifications (Koda 2004). The five properties of the fundamental frequency of initial and repeated calls were acoustically analyzed. The results showed immediate acoustic modifications in the SRS, suggesting exaggerated acoustic features in the repeated coo call. In the SRS, the repeated coo call had a higher fundamental frequency and longer duration than the initial coo call (Fig. 3). In particular, the statistical analysis, which showed improvement from the previous study (Koda 2004), showed significant differences in the maximum and average frequencies and duration between the initial and repeated coo calls. The acoustic analysis of SRSs suggests the possibility that monkeys repeat a modified coo call within a short interval to elicit a vocal response from other group members.

The naturalistic observations confirmed that repeated coo calls, compared to the initial calls, more effectively elicited response coo calls from other group members, which concurs with the previous acoustic analysis (Koda 2004). Vocal transition analysis showed a significant effect of the repeated call on eliciting the response call. To examine the transition of the caller in the vocal exchange, the transitional pattern of the callers during the vocal exchange was analyzed. Figure 4 shows the transitional patterns and probabilities involving the subject macaques and other group members on the basis of the naturalistic observation of the coo exchange. In total, 396 coo calls were recognized from the subject (A1) after a silence of 2.0 s. In the vocal transition pattern, within 2.0 s of A1, 158 calls from the same subject (A2) and 69 calls from other group members (B1) were observed (transitional probabilities: 0.43 and 0.19, respectively; Fig. 4). Within a further 2.0 s, 31 calls from the same subject (A3) and 58 calls from other members (B2) followed A2 (transitional probabilities: 0.21 and 0.39, respectively; Fig. 4). Comparisons of the transitional probability that B2 followed A2 to the probability that B1 followed A1 revealed that the former was significantly greater than the latter. This suggests that repeated coo calls more effectively elicited a vocal response from other members than did initial coo calls. This could be explained by either the enhanced acoustic properties of the repeated coo call.

The subsequent field experiment confirmed that increases in response rates after the repeated coo call were the result of a change in the acoustic properties between the initial and repeated calls (Koda 2004). To examine whether the acoustic properties of the modified coo call are more effective in gaining the attention of other group members, playback experiments were conducted for the seven subject macaques. Figure 5 presents the response rate of the seven subjects, showing a significantly greater response rate for repeated coo stimuli than for initial coo stimuli. The experimental approaches replicated the results of the naturalistic observations. Consequently, the repeated coo call could elicit a vocal response more effectively than the initial coo call.

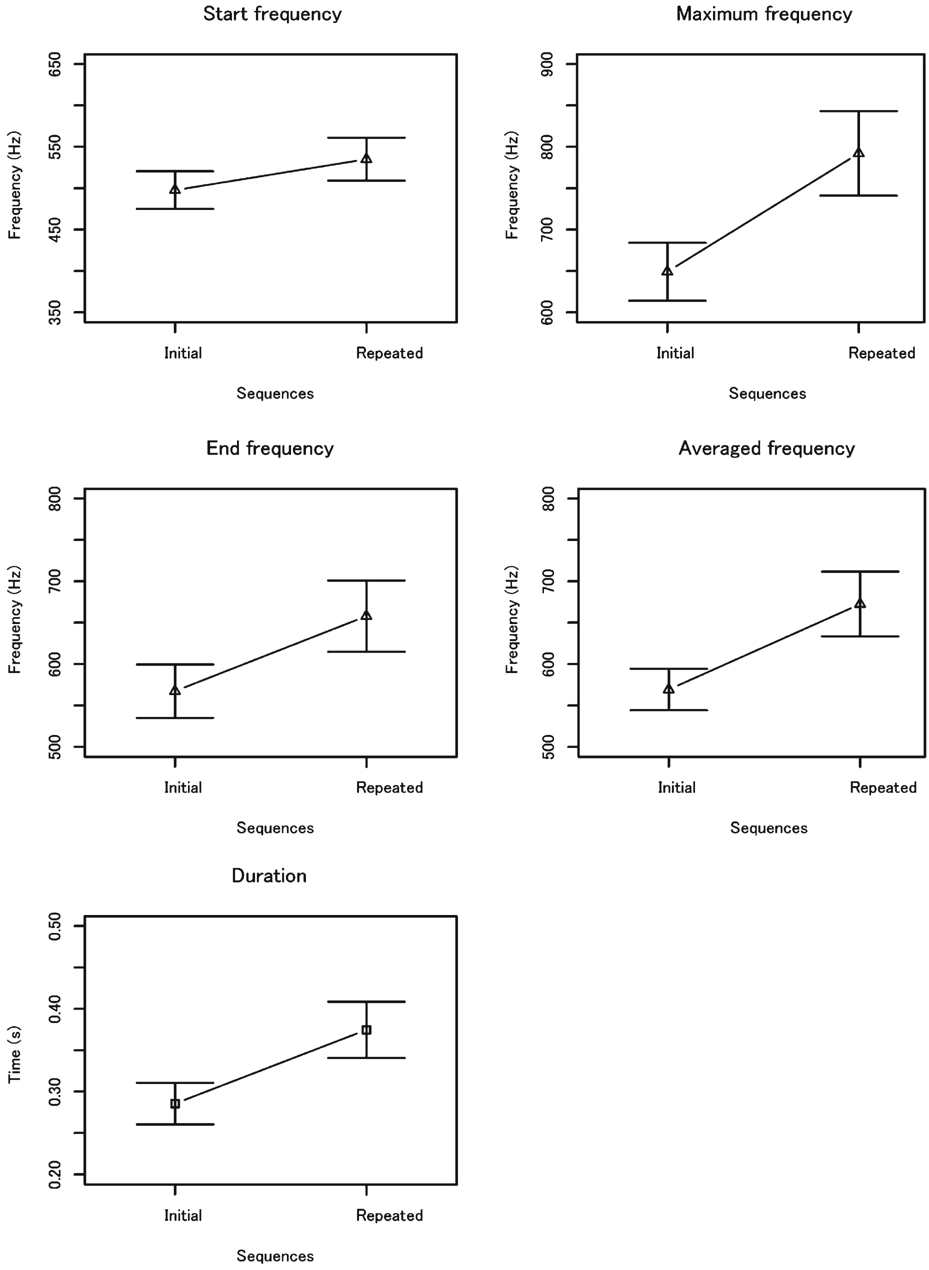


Fig. 3. Mean and standard errors of five acoustic properties of both initial and repeated coo calls. The open triangles represent the mean values of 11 female subjects. The statistical procedure of a mixed-model analysis of variance (ANOVA) with call order (initial call, repeated call) as a fixed factor and individual and sequential utterance of SRS as random factors, showed the significant differences of acoustic properties between the initial and repeated coo calls (start frequency: $F_{1,10.1} = 1.91$, $P = 0.20$; maximum frequency: $F_{1,9.26} = 16.6$, $P = 0.003$; end frequency: $F_{1,6.33} = 6.57$, $P = 0.041$; averaged frequency: $F_{1,10.3} = 11.6$, $P = 0.007$; duration: $F_{1,14.7} = 5.07$, $P = 0.04$, modified from Koda 2004, with permission).

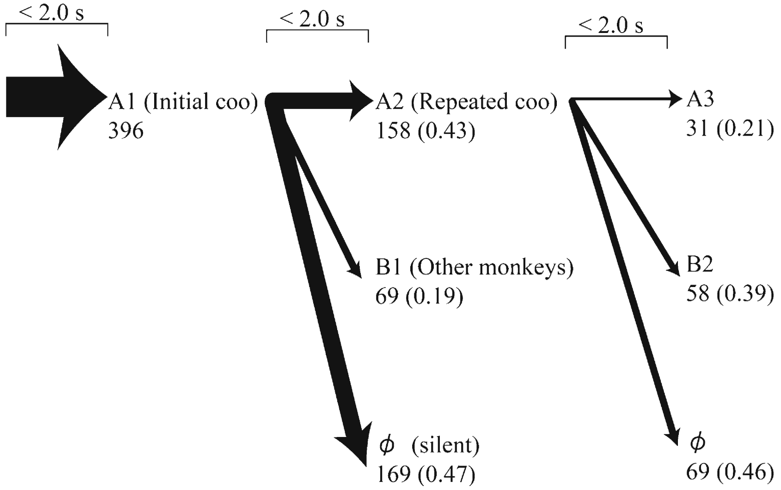


FIG. 4. Vocal transition following a coo call preceded by a silence of 2.0 s. The widths of the arrows represent the frequency of the subsequent events. The transitional probability is shown in brackets. A, B, and phi (*W) represent the calls of the subject, other members, and a silence >2.0 s, respectively (modified from Koda 2004, with permission).

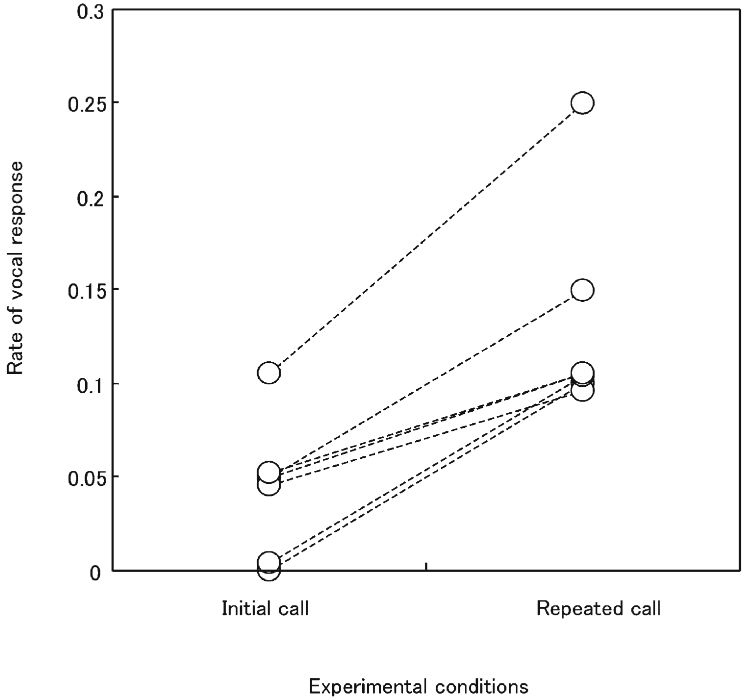


FIG. 5. Results of the playback experiments. Open circles indicate the vocal response rates of each of seven subjects (modified from Koda 2004, with permission).

3 Comparative Analysis of Conversational Communication in Human and Nonhuman Primates

3.1 Accumulating Evidence of Short-term Acoustic Modification in Nonhuman Primates

Monkeys are likely to modify the acoustic structure of a coo call to elicit a vocal response when the initial call attempt fails. The repeated coo call could attract the attention of other members and is more effective in eliciting a response vocalization. This pattern represents context-dependent vocal communication among nonhuman primates.

The acoustic modification is also likely to serve in the efficient maintenance of vocal contact. Although it has not been directly confirmed whether acoustic modification enhances the efficiency of vocal contact in wild groups, previous studies of acoustic modification in laboratory subjects have provided positive evidence. In captive squirrel monkeys, Masataka and Symmes (1986) found that when separated from their group members, subjects prolonged the duration of contact calls in accordance with the separation distance. Moreover, a high-frequency element in this longer call was particularly prolonged. The authors concluded that the acoustic prolongation ensured the efficiency of the vocal contact. In captive common marmosets, Schrader and Todt (1993) revealed a similar phenomenon. The visual and auditory isolation from group members provoked the acoustic modification of contact calls with a prolonged duration and higher modulated pitch. These findings also support the contribution of acoustic modification to communication efficiency. Moreover, recent studies have focused on the prolongation phenomenon. In captive common marmosets, Brumm et al. (2004) clearly confirmed the prolongation of contact calls in accordance with increasing levels of background noise. The contact call prolongation was also observed in captive cotton-top tamarins under the experimental condition of background noise masking (Egnor and Hauser 2006; Egnor et al. 2006). These five studies directly showed short-term acoustic modification, strongly suggesting its contribution to ensure turn-taking in contact communication. These phenomena are difficult to confirm in wild groups. Although the dynamic modification of contact calls has rarely been reported in natural vocal interactions in wild groups, my findings and those of Sugiura (1998) provide positive evidence that acoustic modification ensures efficient turn-taking in vocal communication, at least in wild Japanese macaques.

3.2 Comparison to Motherese in Human Mother–Infant Interactions

Contrary to the controversial issue of vocal plasticity in nonhuman primates, clear evidence exists that acoustic modification in the dynamic interaction between speakers is typical of human conversation. In particular, exaggerations of prosodic features are frequently observed in mother–infant interactions (for review, see Falk 2004). Adult humans modify their speech style into what is generally known as *motherese* when addressing infants and young children. Motherese is a term that has historically referred to the prosodic exaggerations that typically characterize mothers' speech directed toward infants and young children. Specifically, infant-directed (ID) speech cross-linguistically shows that when mothers talk to their infants, the melodic and rhythmic qualities of maternal voices increase (Cooper et al. 1997; Fernald and Simon 1984; Fernald et al. 1989; Grieser and Kuhl 1988; Papousek and Hwang 1991; Shute and Wheldall 1989). ID speech has special acoustic properties such as a high pitch and exaggerated pitch contour. Generally, infants prefer ID speech to adult-directed (AD) speech (Cooper and Aslin 1990; Cooper et al. 1997; Fernald 1985; Pegg et al. 1992; Werker and McLeod 1989). ID speech appears to play a role in the regulation of infants' arousal and attention. On the basis of ethology, the function of ID speech is to gain the attention of the infant, which is likely comparable to the contact call in wild Japanese macaques.

Interestingly, a phenomenon similar to the repeated addressing attempt of wild Japanese macaques was reported in a previous study of human mother–infant interactions (Masataka 1992). When the initial attempts of a mother to attract the attention of her infant failed, she was likely to exaggerate the prosodic features of ID speech compared to the initial attempt of ID speech. Moreover, as the mother repeatedly attempted to address the infant using ID speech, the pitch contour of the ID speech was elevated until the mother gained the infant's attention. Although the human mother–infant interaction is not perfectly homologous to the vocal exchange of nonhuman primates via contact calling, the short-term acoustic modifications of both human mothers and wild Japanese macaques definitely possess common functions to attract the attention of infants or group members. Such context-sensitive patterns of vocal interactions therefore seem to be shared by humans and some nonhuman primate species.

3.3 What Properties Do Human and Nonhuman Primate Communication Share?

The dynamic vocal interactions of both human and nonhuman primates are probably achieved based on some common behavioral properties in their auditory communication. Although studies of wild and captive monkeys have accumulated evidence of greater plasticity in their vocal production than previously considered, the majority of studies do not deny the limitations of primate vocal productions (Egnor and Hauser 2004). It is reasonable to assume that acoustic modification of the addressing vocalization is explained by several fundamental characteristics, as well as by the degree of vocal plasticity.

First, the emotional state is likely to influence acoustic properties in both humans and monkeys. Banse and Scherer (1996) evaluated the acoustic profiles of human speech in vocal emotion expression and found a correlation between the acoustic properties and several expressions of emotion. Similar approaches using multivariate statistical techniques were used in squirrel monkeys (Fichtel et al. 2001). The result showed that increased aversion was paralleled by an increase in frequency range. The phenomenon of pitch exaggeration may be explained by emotion-induced mechanisms, e.g., short-term increases in the arousal levels of Japanese macaques or human mothers. When addressing attempts fail, the arousal level might increase rapidly to elicit a response call, thereby affecting the acoustic properties of the repeated addressing calls.

Second, better detectability of higher frequency sounds has been confirmed in the auditory perception of both human and nonhuman primates (for review, see Kojima 2003). In human mothers, ID speech with prosodic exaggerations is more detectable than AD speech, just as the repeated attempts of coo calls are more detectable than the initial addressing attempt in wild Japanese macaques. Sound detectability captures the attention of the addressed infant or macaque group member. Of particular interest is the infants' preference for ID speech with its melodic and rhythmic qualities. We currently have little evidence of nonhuman primates' preferences for vocalizations with a higher pitch contour or modified prosodic features, although perceptual research has indicated the detectability of such sounds (e.g., Fay 1988). Although the sound preference in nonhuman primates is a challenging topic for future research, the prosodic features of repeated addressing calls in both human mothers and macaques no doubt contribute to gaining the attention of an infant or group member on the basis of sound detectability.

As I described in this chapter, the vocal exchange and addressing communications show shared properties among human and nonhuman primates beyond the boundaries of vocal plasticity. Interestingly, the prosodic exaggerations observed in human mothers is not only restricted to ID speech, but has also been confirmed for caregivers' speech addressing elderly people (Masataka 2002). These results probably show that humans instinctively switch their speech style to a more exaggerated one during dynamic interactions when addressing attempts are strong. Such instinctive vocal modification during the vocal exchange seems to be shared among human and nonhuman primates. Short-term acoustic modifications in wild Japanese macaques may be an evolutionary origin of motherese.

Acknowledgments. I especially appreciate the helpful comments and support of Professor Nobuo Masataka and Dr. Hideki Sugiura. I thank the Yakushima Environment Conservation Center for permission to perform the research, and the field research center of the Primate Research Institute, Kyoto University, for permission to stay at the field station. This work was supported by Grant-in-Aid for Young Scientists (19730461) from Japan Society for the Promotion of Science (JSPS) and a Grant for the Global Centers of Excellence (COE; A06) to Kyoto University from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan.

References

Banse R, Scherer KR (1996) Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70:614–636

Beecher MD, Stoddard PK, Campbell SE, Horning CL (1996) Repertoire matching between neighbouring song sparrows. Animal Behaviour 51:917–923

Biben M, Symmes D, Masataka N (1986) Temporal and structural analysis of affiliative vocal exchanges in squirrel monkeys (*Saimiri sciureus*). Behaviour 98:259–273

Boinski S, Mitchell CL (1992) Ecological and social factors affecting the vocal behavior of adult female squirrel monkeys. Ethology 92:316–330

Bradbury JW, Vehrencamp SL (1998) Principles of animal communication. Sinauer, Sunderland, MA

Brumm H, Voss K, Kollmer I, Todt D (2004) Acoustic communication in noise: Regulation of call characteristics in a New World monkey. Journal of Experimental Biology 207:443–448

Cheney DL, Seyfarth RM, Palombit R (1996) The function and mechanisms underlying baboon "contact" barks. Animal Behaviour 52:507–518

Cleveland J, Snowdon CT (1982) The complex vocal repertoire of the adult cotton-top tamarin (*Saguinus oedipus oedipus*). Zeitschrift Fur Tierpsychologie 58:231–270

Cooper RP, Aslin RN (1990) Preference for infant-directed speech in the 1st month after birth. Child Development 61:1584–1595

Cooper RP, Abraham J, Berman S, Staska M (1997) The development of infants' preference for motherese. Infant Behavior & Development 20:477–488

Crockford C, Herbinger I, Vigilant L, Boesch C (2004) Wild chimpanzees produce group-specific calls: A case for vocal learning? Ethology 110:221–243

Deacon TW (1989) The homologs to human language circuits in monkey brains. American Journal of Physical Anthropology 78:210–211

Egnor SER, Hauser MD (2004) A paradox in the evolution of primate vocal learning. Trends in Neurosciences 27:649–654

Egnor SER, Hauser MD (2006) Noise-induced vocal modulation in cotton-top tamarins (*Saguinus oedipus*). American Journal of Primatology 68:1183–1190

Egnor SER, Iguina CG, Hauser MD (2006) Perturbation of auditory feedback causes systematic perturbation in vocal structure in adult cotton-top tamarins. Journal of Experimental Biology 209:3652–3663

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

Fay R (1988) Hearing in vertebrates: A psychophysical databook. Hill-Fay Associates, Winnetka

Fernald A (1985) 4-month-old infants prefer to listen to motherese. Infant Behavior & Development 8:181–195

Fernald A, Simon T (1984) Expanded intonation contours in mothers speech to newborns. Developmental Psychology 20:104–113

Fernald A, Taeschner T, Dunn J, Papousek M, Deboyssonbardies B, Fukui I (1989) A cross-language study of prosodic modifications in mothers and fathers speech to preverbal infants. Journal of Child Language 16:477–501

Fichtel C, Hammerschmidt K, Jurgens U (2001) On the vocal expression of emotion. A multi-parametric analysis of different states of aversion in the squirrel monkey. Behaviour 138:97–116

Fischer J, Hammerschmidt K, Todt D (1998) Local variation in Barbary macaque shrill barks. Animal Behaviour 56:623–629

Geissmann T (1999) Duet songs of the siamang, *Hylobates syndactylus*: II. Testing the pair-bonding hypothesis during a partner exchange. Behaviour 136:1005–1039

Ghazanfar A (ed) (2003) Primate audition: Ethology and neurobiology. CRC Press, Boca Raton, FL

Gilissen E (2004) Aspects of human language: Where motherese? Behavioral and Brain Sciences 27:514

Goodwin C (1981) Conversational organization: Interaction between speakers and hearers. Academic Press, New York

Green S (1975) Variation of vocal pattern with social situation in the Japanese monkey (*Macaca fuscata*): A field study. In: Rosenblum LA (ed) Primate behavior, vol 4. Academic Press, New York, pp 1–102

Grieser DL, Kuhl PK (1988) Maternal speech to infants in a tonal language—Support for universal prosodic features in motherese. Developmental Psychology 24:14–20

Hammerschmidt K, Newman JD, Champoux M, Suomi SJ (2000) Changes in rhesus macaque "coo" vocalizations during early development. Ethology 106:873–886

Hammerschmidt K, Freudenstein T, Jurgens U (2001) Vocal development in squirrel monkeys. Behaviour 138:1179–1204

Hauser MD (1991) Sources of acoustic variation in rhesus macaque (*Macaca mulatta*) vocalizations. Ethology 89:29–46

Hauser MD (1992) Articulatory and social factors influence the acoustic structure of rhesus monkey vocalizations—A learned mode of production. Journal of the Acoustical Society of America 91:2175–2179

Itani J (1963) Vocal communication of the wild Japanese monkey. Primates 4:11–66

Janik VM (2000) Whistle matching in wild bottlenose dolphins (*Tursiops truncatus*). Science 289:1355–1357

Koda H (2004) Flexibility and context-sensitivity during the vocal exchange of coo calls in wild Japanese macaques (*Macaca fuscata yakui*). Behaviour 141:1279–1296

Kojima S (2003) A search for the origins of human speech: Auditory and vocal function of the chimpanzee. Kyoto University Press, Kyoto

Krebs JR, Ashcroft R, Vanorsdol K (1981) Song matching in the great tit *Parus major* L. Animal Behaviour 29:918–923

Marler P (1977) The evolution of communication. In: Sebeok TA (ed) How animals communicate. Indiana University Press, Bloomington, IN, pp 45–70

Marshall AJ, Wrangham RW, Arcadi AC (1999) Does learning affect the structure of vocalizations in chimpanzees? Animal Behaviour 58:825–830

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

Masataka N (2002) Pitch modification when interacting with elders: Japanese women with and without experience with infants. Journal of Child Language 29:939–951

Masataka N, Biben M (1987) Temporal rules regulating affiliative vocal exchanges of squirrel monkeys. Behaviour 101:311–319

Masataka N, Symmes D (1986) Effect of separation distance on isolation call structure in squirrel monkeys (*Saimiri sciureus*). American Journal of Primatology 10:271–278

McGregor PK, Dabelsteen T, Shepherd M, Pedersen SB (1992) The signal value of matched singing in great tits—Evidence from interactive playback experiments. Animal Behaviour 43:987–998

Miller PJO, Shapiro AD, Tyack PL, Solow AR (2004) Call-type matching in vocal exchanges of free-ranging resident killer whales, *Orcinus orca*. Animal Behaviour 67:1099–1107

Mitani JC, Brandt KL (1994) Social factors influence the acoustic variability in the long-distance calls of male chimpanzees. Ethology 96:233–252

Mitani JC, Hasegawa T, Groslouis J, Marler P, Byrne R (1992) Dialects in wild chimpanzees. American Journal of Primatology 27:233–243

Owren MJ, Dieter JA, Seyfarth RM, Cheney DL (1992) Food calls produced by adult female rhesus (*Macaca mulatta*) and Japanese (*M. fuscata*) macaques, their normally-raised offspring, and offspring cross-fostered between species. Behaviour 120:218–231

Papousek M, Hwang SFC (1991) Tone and intonation in mandarine Babytalk to presyllabic infants—Comparison with registers of adult conversation and foreign-language instruction. Applied Psycholinguistics 12:481–504

Pegg JE, Werker JF, McLeod PJ (1992) Preference for infant-directed over adult-directed speech—Evidence from 7-week-old infants. Infant Behavior & Development 15: 325–345

Rendall D, Rodman PS, Emond RE (1996) Vocal recognition of individuals and kin in free-ranging rhesus monkeys. Animal Behaviour 51:1007–1015

Rendall D, Cheney DL, Seyfarth RM (2000) Proximate factors mediating “contact” calls in adult female baboons (*Papio cynocephalus ursinus*) and their infants. Journal of Comparative Psychology 114:36–46

Schrader L, Todt D (1993) Contact call parameters covary with social context in common marmosets, *Callithrix. j. jacchus*. Animal Behaviour 46:1026–1028

Shute B, Wheldall K (1989) Pitch alterations in British motherese—Some preliminary acoustic data. Journal of Child Language 16:503–512

Sibly RM, Nott HMR, Fletcher DJ (1990) Splitting behavior into bouts. Animal Behaviour 39:63–69

Smuts BB, Cheney DL, Seyfarth RM, Wrangham RW, Struhsaker TT (1987) Primate societies. University of Chicago Press, Chicago

Snowdon CT, Cleveland J (1984) Conversations among pygmy marmosets. American Journal of Primatology 7:15–20

Sugiura H (1993) Temporal and acoustic correlates in vocal exchange of coo calls in Japanese macaques. Behaviour 124:207–225

Sugiura H (1998) Matching of acoustic features during the vocal exchange of coo calls by Japanese macaques. Animal Behaviour 55:673–687

Sugiura H (2001) Vocal exchange of coo calls in Japanese macaques. In: Matsuzawa T (ed) Primate origins of human cognition and behavior. Springer-Verlag Tokyo, Tokyo, pp 135–154

Sugiura H (2007) Adjustment of temporal call usage during vocal exchange of coo calls in Japanese macaques. Ethology 113:528–533

Sugiura H, Tanaka T, Masataka N (2006) Sound transmission in the habitats of Japanese macaques and its possible effect on population differences in coo calls. Behaviour 143:993–1012

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

Werker JF, McLeod PJ (1989) Infant preference for both male and female infant-directed talk—A developmental study of attentional and affective responsiveness. Canadian Journal of Psychology 43:230–246

Winter P (1969) Variability of peep and twit calls in captive squirrel monkeys (*Saimiri sciureus*). Folia Primatologica 10:204–215

6
Vocal Learning in Nonhuman Primates: Importance of Vocal Contexts

Chieko Yamaguchi[1,2] and Akihiro Izumi[2]

1 To What Extent Do Nonhuman Primates Show Vocal Flexibility?

Humans acquire languages indeed naturally. Although newborn infants can not speak anything, they acquire normal speech by hearing adults' conversations without some explicit training. Human infants learn a language which they are exposed to in childhood. Although it is impossible for adults to acquire a new language in the same level of the language which was learned in childhood, adults learn some new languages with considerable efforts.

Are these abilities of acquiring languages unique in humans? From the beginning of the 1900s, researchers have attempted to teach great apes to produce human speech. Furness (1916) reported that an orangutan learned to speak a number of human speech sounds voluntarily. Hayes and Hayes (1951) raised a chimpanzee called Viki in their home as one of their own children. Viki became to understand the spoken English of her caregivers to some extent. Although she could say four words of mama, papa, cup, and up, she was never able to say more than these.

It was difficult to teach spoken languages to great apes, but great apes seem to be able to communicate with humans using sign language. Gardner and Gardner (1969) taught sign language to a young chimpanzee named Washoe, and Washoe became to be able to make many of different signs. Patterson (1978) also reported to succeed teaching signs to a gorilla named Koko. The successful acquisition of signs suggested that apes' limited abilities to produce new vocalizations resulted in the failure to learn spoken language.

Although it seems to be difficult for nonhuman primates to learn producing some new vocalizations, it does not mean their vocalizations lack flexibility completely. Some studies reported that nonhuman primates control their

[1] Department of Behavioral and Brain Sciences, Primate Research Institute, Kyoto University, 41 Kanrin, Inuyama, Aichi 484-8506, and [2] Department of Animal Models for Human Disease, National Institute of Neuroscience, National Center of Neurology and Psychiatry, 4-1-1 Ogawa-Higashi, Kodaira, Tokyo 187-8502, Japan

species-specific vocalizations according to various environmental parameters. One of the most popular effects of modifying vocalizations according to environments is called "Lombard effect". Lombard effect is a phenomenon that a signaler increases the amplitude of its vocalizations in response to an increase in the background noise level. As in humans, such an effect has been shown in macaques (Sinnott et al. 1975), common marmosets (Brumm et al. 2004) and cotton-top tamarins (Egnor and Hauser 2006). Brumm et al. (2004) showed that common marmoset prolonged syllable duration and increased amplitude of their twitter calls in response to increased levels of background noise. Because small voices were not able to be transmitted to other individuals in high level of noise, changing the amplitude of vocalizations in response to levels of background noise seems to be important in vocal communication.

Several studies reported another example that nonhuman primates modify their vocalizations according to environments. There are differences in the acoustic structures of vocalizations elicited by geographically distant populations: like "dialects" in humans. In humans' dialects, not only there are differences in acoustic structures but also completely different words are used between distant populations, and it is thought that some of the dialects coursed by difference of cultures. In nonhuman primates, the acoustic differences were thought to be coursed mainly by environmental factors. Tanaka et al. (2006) showed that coo calls of Japanese monkeys living in two geographically distinct populations (Yakushima and Ohirayama) were acoustically different. Tanaka et al. mentioned that the habitat of the Yakushima group is evergreen forest, and Ohirayama group is dirt or gravelly ground with little vegetation. The results were that Ohirayama population vocalized coo calls in lower frequency than Yakushima population. Sugiura et al. (1999) reported that the Ohirayama habitat has a louder ambient noise than the Yakushima habitat, and lower frequency coo calls were more efficiently transmitted in the Ohirayama habitat. Similar acoustic differences were showed in Japanese monkeys' food calls (Green 1975) and chimpanzees' pant hoots (Crockford et al. 2004; Marshall et al. 1999; Mitani et al. 1992).

There are vocal modifications not only according to environments but also according to conspecifics' vocalizations in communicative situations. Sugiura (1998) examined the coo call plasticity during vocal exchange in wild Japanese monkeys. When the monkeys responded to a preceding coo call from another individual, they matched fundamental frequency of their replies to those of the preceding calls. The monkeys seemed to match acoustic structure to show which calls were responded when theys responded to a preceding call. Koda (2004) found that Japanese monkeys tended to emit calls in high frequency when other monkeys did not respond to the preceding call. He discussed that the higher frequency of the repeated calls were more effective in eliciting vocal responses from conspecifics.

It seems that nonhuman primates modify their vocalizations, and such a modification seems to be effective to communicate with others in various environmental conditions. How nonhuman primates acquire these effective modifications of vocalizations? Humans acquire their languages without explicit training. Do nonhuman primates acquire their abilities from learning?

2 Vocal Learning in Infants

There are three types of vocal learning: 1) vocal comprehension learning; 2) vocal usage learning; 3) vocal production learning (Egnor and Hauser 2004; Janik and Slater 2000). Vocal comprehension learning is to learn appropriate responses to others' vocalizations. Vocal usage learning is to learn which calls are used in which contexts. Vocal production learning is to learn how to emit calls.

There are a lot of evidences for vocal comprehension learning and vocal usage learning in nonhuman primates. For vocal comprehension learning, Fischer et al. (2000) examined the development of infant baboons' responses to two kinds of vocalizations: contact barks and alarm barks. Contact barks are given when the caller is at risk of losing contact with the groupmates, and alarm barks are given when the caller spots predators. The two kinds of vocalizations are similar in acoustic structures, and only vocal tonality is different. Fischer et al. played back the two kinds of vocalizations, and measured the time of looking at the speaker for each vocalization. Infants who were two and half months old did not respond to the calls, whereas four-months-old infants looked at the speaker irrespective of the call types. Six-months-olds looked at the speaker when alarm barks were played back, but they did not respond to contact barks. Fischer et al. discussed that infant baboons learn to attend to barks at first, and later learn to discriminate between different bark types. It suggests that the monkeys learn vocal comprehension gradually in their developing process.

For vocal usage learning, Seyfarth et al. (1980) showed that vervet monkeys acquire usage of vocalizations according to situations in development. Vervet monkeys give three different alarm calls to three different predators: leopards, eagles, snakes. Although infants primarily gave leopard alarms to various mammals, eagle alarms to many birds, and snake alarms to various snakelike objects, they improved predator classifications with experiences. McCowan and Newman (2000) conducted a playback experiment and reported both vocal comprehension and usage learning in squirrel monkeys. They played back chuck calls that were called by individuals of various levels of associations to the subject monkeys. Although adults responded preferentially to vocalizations of familiar group members, infants did not.

Compared to vocal comprehension and usage learning, vocal production learning in nonhuman primates is much less clear. Humans require auditory experiences during an early sensitive period in order to develop normal speech (Oller and Eilers 1988; Stoel-Gammon and Otomo 1986). Although many kinds of studies have been conducted to examine whether nonhuman primates require auditory experiences to develop normal vocal behavior, the results were inconsistent. Newman and Symmes (1974) showed that the vocalizations of isolate-reared rhesus macaques were different from the vocalizations of individuals who were reared with conspecifics. Winter et al. (1973) and Hammerschmidt et al. (2000) reported that there was no difference between the vocalizations of isolate reared and reared with their mothers. In studies that examined effects of deafening on vocal development in infants, Winter et al. (1973) showed that there was no difference of vocalizations between deafened and normal infant. Masataka and Fujita (1989) have reported that infants of cross-fostered rhesus monkeys and Japanese monkeys acquired an acoustic feature of the food coo of their foster mother. However, Owren et al. (1992) showed that the food coo of the two species were not different significantly in acoustic structure.

It seems that monkeys learn vocal comprehension and usage, but there are not enough evidences that monkeys learn vocal production in natural environments. It contrasts with the facts that monkeys control their vocalizations according to various conditions and there are evidences of vocal comprehension and usage learning. Because human infants learn all of vocal comprehension, usage and production, a major gap between humans and monkeys seems to exist in vocal production learning.

3 Conditioning of Vocal Behavior

To examine whether or not nonhuman primates learn vocalizations, there are other methods that is to train vocalizations directly. To train some new behavior, a method called operant conditioning is used. Operant conditioning is a process of changing the emergence rate of one behavior by the results of the behavior. For examples, if an experimenter wants a monkey to press a lever, the experimenter repeats to give a reward when the monkey accidentally presses the lever. The monkey becomes to presses the lever voluntarily to get rewards.

Various species of animals have been conditioned to vocalize operantly. Dogs (Salzinger and Waller 1962), rats (Lal 1967), and mynahs (Hake and Mabry 1979) were successfully conditioned to increase rates of vocal responses. Manabe et al. (1995) successfully conditioned to vocalize different calls according to different stimuli with food rewards in budgerigars. Cats were conditioned to increase not only rates of vocal responses but also vocal duration with food rewards (Molliver 1963). By using intracranial self stimulations, Burnstein and Wolff (1967) trained guinea pigs to vocalize in a frequency range. The guinea pigs increased the rates of vocalizations in the frequency range compared to those out of the range.

Various species of nonhuman primates also have been used for subjects of vocal conditioning: lemurs (Wilson 1975), capuchins (Leander et al. 1972; Myers et al. 1965), macaques (Aitken and Wilson 1979; Sutton et al. 1973; 1978; 1981; Yamaguchi and Myers 1972), gibbons (Haraway et al. 1981; Maples and Haraway 1982) and chimpanzees (Kojima 2003; Randolph and Brooks 1967). Although one study reported failures to achieve conditioning (Yamaguchi and Myers 1972), the other studies reported successfully conditioning of vocal behavior to some extent. Both adults and juveniles were used for subjects in these studies, but there were not clear differences in results between adults and infants.

In most of these studies, monkeys were conditioned to vocalize or not according to some stimuli: vocal usage learning. For examples, monkeys were trained to vocalize when a light was on and to be silence when a light was off (e.g., Aitken and Wilson 1979; Myers et al. 1965). Although there have been many studies that showed successful conditioning, it is not to say that all subjects were conditioned to vocalize strongly. As mentioned, one study failed to show vocal conditioning in monkeys (Yamaguchi and Myers 1972). Leander et al. (1972) and Wilson (1975) reported low rates of responding by the monkeys, and Myers et al. (1965) and Aitken and Wilson (1979) reported that not all monkeys were able to achieve conditioning.

Sutton et al. (1973) successfully trained rhesus monkeys to vocalize longer and louder. Although the results seem to be a rare evidence of vocal production learning in monkeys, Janik and Slater (1997) suggested the changes in overall call duration and amplitude do not represent evidence for vocal production learning. They discussed such vocal changes do not require any changes in the setting of the sound production organ, but only a longer or stronger expiration phases.

Can we conclude the nonhuman primates learned to emit vocalizations voluntarily in the previous vocal-operant studies which reported successful conditioning? Although it seems difficult for nonhuman primates to show vocal conditioning, it is not that nonhuman primates can not control their communication behavior at all. Louboungou and Anderson (1987) attempted to condition yawning, scratching, and lip protrusion with food rewards in pigtail macaques. Although lip protrusion was difficult, scratching and yawning was conditioned quickly. Scratching behavior was also successfully conditioned in lemurs (Anderson et al. 1990), long-tailed macaques (Mitchell and Anderson 1993) and vervet monkeys (Iversen et al. 1984). Anderson and Wunderlich (1988) reported that yawning behavior was also successfully conditioned in Tonkean macaques with food rewards. Izumi et al. (2001) reported that rhesus monkeys were successfully conditioned to express three facial actions (tongue protrusion, mouth opening and mouth distortion) in response to arbitrary visual cues. They discussed that monkeys can control facial actions in the absence of social context if the monkeys are appropriately trained. It is interesting that although vocalizations seem to be difficult for monkeys to control, mouth movements do not. Vocal conditioning in the previous studies might be different from other operant behavior in nonhuman primates, and it is also different from vocal learning in humans.

4 Importance of Social Contexts in Vocal Learning

Why is it difficult to train vocal behavior in nonhuman primates? One possible explanation is that nonhuman primates can not vocalize independently from its contexts. For examples, Kojima (2003) trained a chimpanzee to vocalize differentially according to two kinds of rewards: vocalize “o” for milk and “a” for banana. Although he reported that the chimpanzee gradually learned to vocalize “a” for banana, he did not report about vocalizing “o” for milk. The results might be plausible if chimpanzees in general tend to vocalize like “a” when they find food. The difficulty of training to vocalize “o” for milk might be caused by that chimpanzees do not tend to vocalize like “o” when they find milk. The failure to train nonhuman primates to vocalize in the previous studies might be caused by mismatching between trained vocalizations and contexts.

Human vocal learning seems to be free from contexts. For example, humans are possible to acquire new language to some extent by hearing radio programs (with certain effort). Compared to the vocal learning in humans, those in nonhuman primates seem to be highly restricted in particular contexts. Pierce (1985) pointed out social rewards seem to facilitate vocal conditioning of nonhumans. Randolph and Brooks (1967) showed successful conditioning to vocalize with play rewards in a chimpanzee. Haraway et al. (1981) and Maples and Haraway (1982) successfully trained gibbons to vocalize longer period per session with rewards of taped conspecifics' vocalizations. If vocal learning were restricted in particular contexts, it would be difficult to train chimpanzees to produce play vocalizations to get food rewards. In the previous vocal-conditioning studies, nonhuman primates might not emit vocalizations voluntarily to get rewards. Instead, vocalizations seem to be elicited by some stimuli which made the subjects anticipate rewards.

Different stimuli can elicit vocalizations that are acoustically distinct. Two studies showed that nonhuman primates spontaneously differentiated vocalizations according to different contexts without training vocalizations directly. Taglialatela et al. (2003) showed that an adult male bonobo named Kanzi produced distinct vocalizations with distinct contexts. Kanzi vocalized spontaneously when he used lexigram, gestured to object and responded to yes-no questions. The vocalizations when Kanzi behaved to indicate banana, grape, juice and yes were different from each other. Hihara et al. (2003) trained a Japanese monkey to use a rake-shaped tool to retrieve distant food. The trained monkey began to vocalize coo calls spontaneously in tool-using context. Hihara et al. then trained the monkey to vocalize to request food or the tool, and reinforced whatever kind of coo calls. The acoustic analysis revealed the coo calls were differentiated according to situations when the monkey requested for either food or the tool. The subjects in these studies seem to modify their vocalizations slightly in response to the experimental contexts.

What kinds of contexts affect nonhumans' vocalizations strongly? Because vocal communication is what to exchange some information with other individuals, it is likely that social contexts are important in vocal communication. Some reports showed that social contexts, especially other individuals' states, practically affect vocalizations. At a feeding context, captive nonhuman primates are known to emit vocalizations that were called food calls. Food calls refer to vocalizations uttered by an animal upon encountering a food source, and the calls appear to be attractive to conspecifics (Dittus 1984; Elowson et al. 1991). Some studies showed that nonhuman primates modify their food calls according to various conditions. Hauser et al. (1993) examined effects of food quantity and divisibility on food calls in chimpanzees. The chimpanzees vocalized more frequently when they found large amount of food. Hauser et al. also presented divisible and nondivisible food of the same amount, and the chimpanzees vocalized more frequently when they found divisible food. In New World monkeys, effects of audience status on food calls were examined. When red-bellied tamarins discovered food, they uttered calls at a higher rate when they were isolated from the rest of the group than in the presence of groupmates (Caine et al. 1995). Vitale et al. (2003) showed that common marmosets vocalized more frequently when the groupmates were separated in different cage rooms than when the groupmates were separated by a wire mesh (i.e., they could see each other). If one individual discovered food when it was with groupmates, it is not necessary to emit food calls because the groupmates can discover the same food by themselves. These results suggested that monkeys might modify their vocal behavior according to whether or not the conspesifics should receive the information.

Monkeys can control their vocal behavior not only in natural communicative situation with conspecifics but also in experimental contexts with humans. Yamaguchi and Izumi (2006) showed that human experimenters' states affect vocalizations of Japanese monkeys in the food begging situations. In the experiment, an experimenter who had a food bowl moved toward and away from a hungry monkey. The monkeys vocalized more frequently when the experimenter moved away rather than towards them. The monkeys also vocalized more frequently when the experimenter stood in front of the monkeys and faced away rather than faced towards. These results suggested that the monkeys vocalized more frequently when the situation changed to where the monkeys were not likely to get food from the experimenter. It seems that the monkeys recognize the attentional states of others by body orientations and modify their vocal behavior accordingly.

Nonhuman primates seem to control their vocal behavior flexibly with no explicit training in social contexts. As well as humans, nonhuman primates can modify their vocalizations to communicate effectively with other individuals who should receive information. Although many previous studies in laboratories did not take care of the effects of social contexts on vocal learning, such a context might be a key factor to investigate abilities of vocal learning in nonhuman primates.

References

Aitken PG, Wilson WA (1979) Discriminative vocal conditioning in rhesus monkeys: Evidence for volitional control? Brain and Language 8:227–240

Anderson JR, Wunderlich D (1988) Food-reinforced yawning in *Macaca tonkeana*. American Journal of Primatology 16:165–169

Anderson JR, Fritsch C, Favre B (1990) Operant conditioning of scratching in lemurs. Primates 31:611–615

Brumm H, Voss K, Kollmer I, Todt D (2004) Acoustic communication in noise: Regulation of call characteristics in a New World monkey. Journal of Experimental Biology 207:443–448

Burnstein DD, Wolff PC (1967) Vocal conditioning in the guinea pig. Psychonomic Science 8:39–40

Caine NG, Addington RL, Windfelder TL (1995) Factors affecting the rates of food calls given by red-belled tamarins. Animal Behaviour 50:53–60

Crockford C, Herbinger I, Vigilant L, Boesch C (2004) Wild chimpanzees produce group-specific calls: A case for vocal learning? Ethology 110:221–243

Dittus WPJ (1984) Toque macaque food calls: Semantic communication concerning food distribution in the environment. Animal Behaviour 32:470–477

Egnor SER, Hauser MD (2004) A paradox in the evolution of primate vocal learning. Trends in Neurosciences 27:649–654

Egnor SER, Hauser MD (2006) Noise-induced vocal modulation in cotton-top tamarins (*Saguinus oedipus*). American Journal of Primatology 68:1183–1190

Elowson AM, Tannenbaum P, Snowdon CT (1991) Food-associated calls correlate with food preferences in cotton-top tamarins. Animal Behaviour 42:931–937

Fischer J, Cheney DL, Seyfarth RM (2000) Development of infant baboons' responses to graded bark variants. Proceedings of the Royal Society of London. Series B, containing papers of a biological character. Royal Society (Great Britain) 267:2317–2321.

Furness WH (1916) Observations on the mentality of chimpanzees and orang-utans. Proceedings of American Philosophical Society 55:281–290

Gardner RA, Gardner BT (1969) Teaching sign language to a chimpanzee. Science 165:664–672

Green S (1975) Dialects in Japanese monkeys: Vocal learning and cultural transmission of locale-specific vocal behavior? Zeitschrift für Tierpsychologie 38:301–314

Hake DF, Mabry J (1979) Operant and nonoperant vocal responding in the mynah: Complex schedule control and deprivation-induced responding. Journal of the Experimental Analysis of Behavior 32:305–321

Hammerschmidt K, Newman JD, Champoux M, Suomi S (2000) Changes in rhesus macaque "coo" vocalization during early development. Ethology 106:873–886

Haraway MM, Maples EG, Tolson S (1981) Taped vocalization as a reinforcer of vocal behavior in a siamang gibbon (*Symphalangus syndactylus*). Psychological Reports 49:995–999

Hauser MD, Teixidor P, Field L, Flaherty R (1993) Food-elicited calls in chimpanzees: Effects of food quantity and divisibility. Animal Behaviour 45:817–819

Hayes KJ, Hayes C (1951) The intellectual development of a home-raised chimpanzee. Proceedings of the American Philosophical Society 95:105–109

Hihara S, Yamada H, Iriki A, Okanoya K (2003) Spontaneous vocal differentiation of coo-calls for tools and food in Japanese monkeys. Neuroscience Research 45:383–389

Iversen JH, Ragnarsdottir GA, Randrup KI (1984) Operant conditioning of autogrooming in vervet monkeys (*Cercopithecus aethiops*). Journal of the Experimental Analysis of Behavior 42:171–189

Izumi A, Kuraoka K, Kojima S, Nakamura K (2001) Visually guided facial actions in rhesus monkeys. Cognitive, Affective & Behavioral Neuroscience 1:266–269

Janik VM, Slater PJB (1997) Vocal learning in mammals. Advances in the Study of Behavior 26:59–98

Janik VM, Slater PJB (2000) The different roles of social learning in vocal communication. Animal Behaviour 60:1–11

Koda H (2004) Flexibility and context-sensitivity during the vocal exchange of coo calls in wild Japanese macaques (*Macaca fuscata yakui*). Behaviour 141:1279–1296

Kojima S (2003) A search for the origins of human speech: Auditory and vocal functions of the chimpanzee. Kyoto University Press/Trans Pacific Press, Kyoto/Melbourne, pp 125–127

Lal H (1967) Operant control of vocal responding in rats. Psychonomic Science 8:35–36

Leander JD, Milan MA, Jaspper KB, Heaton KL (1972) Schedule control of the vocal behavior of cebus monkeys. Journal of the Experimental Analysis of Behavior 17:229–235

Louboungou M, Anderson JR (1987) Yawning, scratching, and proturuded lips: Differential conditionability of natural acts in pigtail monkeys (*Macaca nemestrina*). Primates 28:367–375

Manabe K, Kawashima T, Staddon JER (1995) Differential vocalization in budgerigars: Towards an experimental analysis of naming. Journal of the Experimental Analysis of Behavior 63:111–126

Maples EG, Haraway MM (1982) Taped vocalization as a reinforcer of vocal behavior in a female agile gibbon (*Hylobates agilis*). Psychological Reports 51:95–98

Marshall AJ, Wrangham RW, Arcadi AC (1999) Does learning affect the structure of vocalizations in chimpanzees? Animal Behaviour 58:825–830

Masataka N, Fujita K (1989) Vocal learning of Japanese and rhesus monkeys. Behaviour 109:191–199

McCowan B, Newman JD (2000) The role of learning in chuck call recognition by squirrel monkeys (*Saimuri sciureus*). Behaviour 137:279–300

Mitani JC, Hasegawa T, Gros-louis J, Marler P, Byrne R (1992) Dialects in wild chimpanzees? American Journal of Primatology 27:233–243

Mitchell RW, Anderson JR (1993) Discrimination learning of scratching, but failure to obtain imitation and self-recognition in a long-tailed macaque. Primates 34:301–309

Molliver ME (1963) Operant control of vocal behavior in the cat. Journal of the Experimental Analysis of Behavior 6:197–202

Myers SA, Horel JA, Pennypacker HS (1965) Operant control of vocal behavior in the monkey. *Cebus albifrons*. Psychonomic Science 3:389–390

Newman JD, Symmes D (1974) Vocal pathology in socially deprived monkeys. Developmental Psychobiology 7:351–358

Oller DK, Eilers RE (1988) The role of audition in infant babbling. Child Development 59:441–449

Owren MJ, Dieter JA, Seyfarth RM, Cheney DL (1992) 'Food' calls produced by adult female rhesus (*Macaca mulatta*) and Japanese (*M. fuscata*) macaques, their normall-raised offspring, and offspring cross-fotered between species. Behaviour 120:218–231

Patterson F (1978) Conversations with a gorilla. National Geographic 154:438–465

Pierce JD (1985) A review of attempts to condition operantly alloprimate vocalization. Primates 26:202–213

Randolph MC, Brooks BA (1967) Conditioning of a vocal response in a chimpanzee through social reinforcement. Folia Primatolica 5:70–79

Salzinger K, Waller MB (1962) The operant control of vocalization in the dog. Journal of the Experimental Analysis of Behavior 5:383–389

Seyfarth RM, Cheney DL, Marler P (1980) Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. Science 210:801–803

Sinnott J, Stebbins WC, Moody DB (1975) Regulation of voice amplitude by the monkey. Journal of the Acoustical Society of America 58:412–414

Stoel-Gammon C, Otomo K (1986) Babbling development of hearing-impaired and normally hearing subjects. Journal of Speech and Hearing Disorders 51:33–41

Sugiura H (1998) Matching of acoustic features during the vocal exchange of coo calls by Japanese macaques. Animal Behaviour 55:673–687

Sugiura H, Tanaka T, Masataka N (1999) Sound transmission in the habitats of Japanese macaques and its effect on populational differences in coo calls. Journal of the Acoustical Society of Japan 55:679–687 (in Japanese)

Sutton D, Larson C, Taylor EM, Lindeman RC (1973) Vocalization in rhesus monkeys: Conditionability. Brain Research 52:225–231

Sutton D, Samson HH, Larson CR (1978). Brain mechanisms in learned phonation of *Macaca mulatta*. In: Chivers DJ, Herbert J (eds) Recent advances in primatology, vol 1: Behavior. Academic Press, London, pp 769–784

Sutton D, Trachy RE, Lindeman RC (1981) Vocal and nonvocal discriminative performance in monkeys. Brain and Language 14:93–105

Taglialatela JP, Savage-Rumbaugh S, Baker LA (2003) Vocal production by a language-competent *Pan paniscus*. International Journal of Primatology 24:1–17

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

Vitale A, Zanzoni M, Queyras A, Chiarotti F (2003) Degree of social contact affects the emission of food calls in the common marmoset (*Callithrix jacchus*). American Journal of Primatology 59:21–28

Wilson WA (1975) Discriminative conditioning of vocalizations in *Lemur catta*. Animal Behaviour 23:432–436

Winter P, Handley P, Ploog D, Schott D (1973) Ontogeny of squirrel monkey calls under normal conditions and under acoustic isolation. Behaviour 138:1179–1204

Yamaguchi C, Izumi A (2006) Effect of others' attentional states on vocalizations in Japanese monkeys (*Macaca fuscata*). Behavioural Processes 73:285–289

Yamaguchi S, Myers RE (1972) Failure of discriminative vocal conditioning in rhesus monkey. Brain Research 37:109–114

7
The Ontogeny and Phylogeny of Bimodal Primate Vocal Communication

Asif A. Ghazanfar[1] and David J. Lewkowicz[2]

1 Introduction

The two primary channels of social communication in primate species are the face and the voice and there is no doubt that each alone can provide perceptually, cognitively and socially meaningful information. When combined, however, the communicative value of the information conveyed through each of these two channels can be greatly enhanced as many stimulus attributes in each channel are invariant and redundant. To be specific, the sender of a communicative signal usually provides the recipient with spatially synchronous facial and vocal signals that also correspond in terms of their duration, tempo, and rhythmic patterning and that usually convey the same linguistic information, affective meaning, and intentions. It is probably precisely because of the fact that the the multisensory signals specifying a vocalizing face are so much more salient than unisensory signals that the primate nervous system has evolved a variety of multisensory integration mechanisms (Ghazanfar and Schroeder 2006; Stein and Meredith 1993). Such mechanisms enable organisms to take full advantage of this increased salience and, in the process, provide for more effective detection, learning, discrimination, and interpretation of perceptual input (Marks 1978; Rowe 1999).

To be comprehensive, an explanation of multisensory communication should should not only involve a consideration of the proximal perceptual and neural mechanisms underlying multisensory integration but also of the more distal developmental and evolutionary processes that contribute to its emergence. This systems perspective is based on the fact that all behavioral traits operate at proximal as well as distal time scales and, thus, that a full understanding of the mechanisms underlying various behavioral capacities must, by necessity, involve explanations at all three levels of organization(Gottlieb 1992; Schneirla 1949). Such a systems perspective has a long tradition in biology. Both Charles Darwin (1871 #4149) and Thomas Huxley (1863) leaned heavily on the facts of embryol-

[1] Neuroscience Institute, Department of Psychology, Princeton University, Princeton, NJ 08540, and [2] Department of Psychology, 777 Glades Road, Florida Atlantic University, Boca Raton, FL 33431, USA

ogy to connect man to the animal kingdom and for indisputable evidence of evolution. Adopting such a systems perspective, in this review of multisensory communication in primates, we will discuss the results from studies investigating commnication at all three levels of organization. For ease of organization, we consider the proximal mechanisms underlying integration of multisensory communication signals throughout the chapter and examine the ontogeny of multisensory perception in human infants first, followed by an examination of the phylogeny of such capacities by reviewing studies with nonhuman primates (primarily macaque monkeys) next, and end with an examination of the neural mechanisms underlying the integration of multisensory communicative signals.

2 The Development of Multisensory Perception & the Effects of Early Experience

The pervasive nature of multisensory interaction and integration has naturally led investigators to ask about the developmental origins of this fundamental process and the last three decades have witnessed great progress in our understanding of when and how various multisensory perceptual abilities emerge during early development, both in humans and animals (Bushnell 1994; Lewkowicz 1994; 2000a; 2002; Lickliter 2000; Walker-Andrews et al. 1997; Wallace et al. 2004). Most of these studies have either explicitly or implicitly been driven by one of two theoretical views. The first, usually referred to as the differentiation view, assumes that basic multisensory perceptual abilities are present at birth and that they become increasingly differentiated and refined over age (Gibson 1969; Gibson 1984). The second, usually referred to as the integration view, assumes that such abilities are not present at birth and that they only emerge gradually during the first years of life as a result of the child's active exploration of the world (Birch and Lefford 1963; 1967; Piaget 1952).

When discussing multisensory perception it is important to note that this is not a unitary phenomenon. Rather, it consists of several distinct processes that include: 1) the ability to perceive equivalent stimulus attributes in different modalities (e.g., duration, tempo, or rhythm) and is often referred to as amodal perception, 2) the ability to associate concurrent modality-specific stimulus attributes (e.g., a shrieking vocalization and the baring of teeth), and 3) non-specific interactions such as those examplified by the McGurk effect (McGurk and MacDonald 1976) where stimulation in one modality changes our perception of stimuli in another modality.

2.1 General Features of the Development of Multisensory Perception

Most of the empirical evidence on the development of multisensory perception comes from studies investigating human infants' ability to perceive equivalent multisensory relations. In general, this evidence has been consistent with the differentiation view in showing that a variety of basic multisensory perceptual abilities appear early in infancy and that as infants grow these abilities change and improve in significant ways (Lewkowicz 1994; 2000a; 2002; Lickliter 2000; Walker-Andrews 1997). More specifically, the evidence has shown that the ability to perceive lower-order multisensory relations emerges first, during the first months of life, and that the ability to perceive more complex multisensory relations emerges during the later months of the first year of life. For example, infants as young as 3 weeks of age can perceive audio-visual relations based on intensity (Lewkowicz and Turkewitz 1980) and infants as young as 2 months of life can perceive multisensory relations based on temporal synchrony relations and can do so regardless of whether such relations are inherent in simple audiovisual events consisting of simple objects and their sounds (Lewkowicz 1986; 1992a; 1992b; 1996) or in complex ones consisting of dynamic and vocalizing faces (Lewkowicz 2000b; 2003). In contrast, it is not until 6 months of age that infants begin to demonstrate the ability to perceive duration/synchrony-based multisensory equivalence (Lewkowicz 1986) and even 8-month-old infants do not exhibit evidence that they can perceive tempo-based multisensory relations (Lewkowicz 1992a). Moreover, it is not until the latter half of the first year of life that infants can perceive higher-order multisensory relations based on such attributes as affect (Walker-Andrews 1986) and gender (Patterson and Werker 2002). It is also not until the latter part of the first year of life that infants become capable of learning arbitrary multisensory associations (Reardon and Bushnell 1988) and of responding to auditory and visual cues cues in a spatial localization task in an integral fashion (Neil et al. 2006). Taken together, this body of findings shows that many of the basic mechanisms that permit infants to integrate auditory and visual sources of information emerge early in life and that those mechanisms change in important ways during the first year of life (Lewkowicz 2002).

Of particular interest in the context of the current chapter is the question of whether and when the ability to integrate the visual and vocal attributes of conspecifics emerges during early development. Like the evidence on the development of responsiveness to basic multisensory relations, the evidence on the development of responsiveness to face-voice relations indicates that the ability to respond to some simple types of relations also emerges relatively early in life. For example, findings indicate that infants as young as 2 months of age are sensitive to simple face-voice relations in that they can perceive the audio-visual equivalence of isolated visible and audible vowels (Kuhl and Meltzoff 1982; Patterson and Werker 2002; 2003). Responsiveness to more complex face-voice relations, such as those specifying affect emerges at 7 months of age. Interestingly, however, this later emergence appears to be true only when the affect is conveyed by strangers' faces and voices (Walker-Andrews 1986). When the affect is conveyed by the infant's own mother, infants as young as 3.5 months of age exhibit evidence of multisensory integration. This finding indicates that experience plays an important role in the development of multisensory perception and raises important questions regarding the general influence of experience in the development of multisensory perception. Considerations of this question can be greatly faciliated by a consideration of the general effects of experience in perceptual development.

2.2 Developmental Broadening

Conventional theories see development as beginning with a set of narrow capacities that gradually broaden in scope as a function of growth, experience, and structural and functional differentiation (Gibson 1969; Gottlieb 1996; Piaget 1952; Werner 1973). Piaget's (1952) theory of cognitive development best exemplifies this view. According to Piaget, we begin life with a set of rudimentary sensorimotor abilities that gradually, over developmental time, become transformed into sophisticated symbolic representational and logico-deductive skills. In other words, the developmental pattern is from simple poorly differentiated information processing skills to highly refined ones that enable us to process, represent, and remember complex information in a highly sophisticated and efficient manner. At the perceptual level, this broadening view sees development as a process of increasing differentiation and specificity for the detection of new stimulus properties, patterns, and distinctive features (Gibson 1969; Werner 1973).

The developmental broadening view represented by most developmental theories is supported by a wealth of empirical evidence. To cite just a few examples, newborn infants have very poor visual acuity and spatial resolution skills, cannot discriminate between different faces, cannot perceive the affect that faces or voices convey, do not understand that objects are bounded and have an independent existence, can hear speech but cannot segment it into its meaningful components, cannot link specific speech sounds to the objects that they represent, cannot understand the meanings inherent in the temporal structure of events, do not perceive depth nor have a fear of heights, and cannot self-locomote. As infants grow, however, these various skills begin to emerge and continually improve over time. A good case-in-point is the development of face perception. At birth, infants exhibit weak and relatively unstable preferences for faces (Morton and Johnson 1991; Nelson 2001), by 2 months of age these preferences become more stable, and by six to seven months of age infants begin to respond to facial affect (Ludemann and Nelson 1988) and begin to categorize faces on the basis of gender (Cohen and Strauss 1979). Another good case-in-point is language development. Although newborn infants possess rudimentary auditory perceptual abilities that enable them to recognize their mother's voice (DeCasper and Fifer 1980) and to distinguish different languages on the basis of their rhythmical attributes (Nazzi et al. 1998), they do not have a functional lexicon. By 1 month of age, infants begin to exibit basic speech perception abilities that enable them to perceive sounds in a language-relevant way (Eimas et al. 1971) and over the next few months quickly acquire various phonetic discrimination abilities that permit them to segment speech into linguistically meaningful components (Jusczyk 1997), extract the statistical properties of speech (Saffran et al. 1996), and learn simple syntactic rules (Marcus et al. 1999). By the second year of life they begin to recognize words and their meanings and begin to link words with their referents (Fernald et al. 1998). Interestingly, as noted above, the pattern of extant findings on the development of multisensory perceptual skills is generally consistent with the developmental broadening view in showing that as infants grow, their multisensory perceptual abilities improve.

2.3 Developmental Narrowing

Although most theories of behavioral development either ignore developmental narrowing processes altogether or assign relatively little importance to them, a class of what might be referred to as “psychobiological theories of development”, do acknowledge their importance through the concept of behavioral canalization. This concept was first introduced into the psychological literature by Holt (1931) to account for the emergence of organized motor activity patterns out of the initially diffuse patterns seen during fetal development. According to Holt, the initially diffuse sensorimotor activity observed in early fetal life becomes canalized into organized motor patterns through the process of behavioral conditioning. Later, Kuo (1976) broadened Holt’s limited concept of canalization by proposing that narrowing of behavioral potential was not merely the result of the individual’s history of reinforcement but that the individual’s entire developmental history, context, and experience play a crucial role in it. It was Gottlieb (1991), however, who provided the first empirical proof of the importance of narrowing in early development in his studies of the development of species identification in ducks. He showed that the development of mallard hatchlings’ socially affiliative responses toward their conspecifics is determined by exposure to their own embryonic vocalizations and, critically, that this experience helps to canalize their subsequent responsiveness. Specifically prior to hatching, embryos vocalize and Gottlieb showed that as they do so they actually learn some of the critical features of their species-specific call and in the process learn *not to respond* to the social signals of other species. In other words, experience with their own embryonic vocalizations narrows the embryos’ initially broadly tuned auditory sensitivity. Once this process is completed, the hatchlings end up being sensitive to the acoustic features of the calls of their own, but not to the calls of other, species.

Similar to psychobiological theories of behavioral development, dynamic systems theories of development (Lewis 2000; Thelen and Smith 1994) have called attention to the importance of regressive developmental processes in their account of the development of motor skills. These theories assume that the degrees of freedom that define the critical parameters that control various motor skills are reduced during motor learning and development. For example, when infants are first learning to walk, the many parts of the motor system are free to assemble into many functional patterns and are free to do so in many different ways (i.e., the degrees of freedom are many). As they begin to move and interact with the physical substrate on which they are trying to locomote, however, the various subparts of their motor system begin to cooperate with one another and begin to assemble into stable and efficient patterns of motor action. As they do so, the functionally useful patterns are selected from the initially many possible ones through a reduction in the degrees of freedom underlying the various subsystems that participate in the control of locomotion. Once such patterns are selected, the system achieves a stable attractor state. In other words, efficient walking emerges from a diffusely organized system that becomes canalized through experience into a more efficient but, at the same time, less plastic system.

Despite the fact that developmental narrowing processes have to some extent been recognized in some developmental quarters, there has not been widespread recognition of the importance of such processes for understanding the eventual products of behavioral development. This situation has, however, been slowly changing as a new and steadily growing body of empirical evidence has begun to provide some convincing proof that a number of human perceptual functions undergo developmental narrowing in early life and that such narrowing is part-and-parcel of the eventual development of species-specific patterns of perceptual expertise. This body of evidence consists of findings from studies of speech and face perception and shows that initially in development some perceptual abilities are broadly tuned and that as development proceeds these abilities become narrower in scope.

2.3.1 Developmental Narrowing in Speech Perception

The original studies that provided the first evidence of developmental narrowing effects in human infants showed that initially present sensitivity to nonnative speech contrasts declines during the first year of life as a function of experience. Specifically, these studies have shown that young infants can perceive both native and nonnative phonetic contrasts but that by the time infants reach the end of the first year of life they can no longer perceive nonnative contrasts. Some of the first, though, indirect evidence of this kind of narrowing came from a study by Streeter (1976). Building on the earlier landmark findings of Eimas and colleagues (Eimas et al. 1971) showing that 1- and 4-month-old infants can perceive a phonologically relevant voice-onset-time contrast (i.e., the distinction between a /ba/ and a /pa/), Streeter showed that 2-month-old Kikuyu infants learning the Kikuyu language can discriminate the voice-onset-time contrast even though this contrast is not used in their native language. Consistent with Streeter's findings, Aslin et al. (1981) found that English-learning 6–12-month-old infants can discriminate a phonologically irrelevant voice-onset-time contrast as well as the phonologically relevant one identified by Eimas et al.

It was Werker and Tees (1984), however, who provided the first direct evidence of developmental narrowing. These investigators demonstrated that 6-to-8-month-old English-learning infants can discriminate nonnative consonants, such as the Hindi retroflex /Da/ and the dental /da/ and the Thompson glottalized velar /k'i/ versus the uvular /q'i/, but that 10-to-12 month-old infants do not discriminate them. Based on these findings, Werker and Tees concluded that the decline in responsiveness to nonnative phonetic contrasts is due to language-specific experience that provides infants with continuing experience with native consonant contrasts and none with the nonnative ones. Subsequent cross-linguistic consonant discrimination studies have provided additional evidence of this type of developmental narrowing. For example, Best et al. (1995) tested 6-to-8 month-old and 10-to-22 month-old English-learning American infants' ability to discriminate native English and Nthlakampx velar vs. uvular ejectives as well as Zulu click sounds. They found that younger infants discriminated the English and Nthlakampx contrasts, older infants only discriminated the English contrasts, and that both age groups discriminated the Zulu click sounds. These authors concluded that the decline in response to Nthlakampx contrasts reflected similar developmental narrowing effects found by Werker and Tees (1984) and that equivalent levels of responsiveness to the Zulu clicks was due to the failure of such clicks to map onto English phonology. Finally, studies have shown that developmental narrowing is not confined to consonants. Kuhl et al. (1992) have shown that responsiveness to nonnative vowels also narrows during infancy and that it actually occurs earlier (by 6 months of age) than in responsiveness to consonants.

2.3.2 Developmental Narrowing in Face Perception

Perceptual narrowing in human infants is not limited to the auditory modality. Studies have shown that infant responsiveness to native and nonnative faces follows a pattern of decline similar to that found in speech perception studies. These studies have found that human as well as monkey adults are better at recognizing faces from their own species than from other species (Dufour et al. 2006; Pascalis and Bachevalier 1998). Once again, the native-nonnative response difference observed in adults turns out to be the result of experience-dependent narrowing processes that occur during infancy. Direct evidence of this fact was first provided by Pascalis et al. (2002) who investigated discrimination of human and monkey faces in 6 and 9-month-old infants. These researchers found that 6-month-old infants discriminated human and monkey faces but that 9-month-old infants only discriminated human faces. In a subsequent study, Pascalis et al. (2005) replicated these findings and, like Kuhl et al. (2003) who showed that responsiveness to nonnative speech contrasts can be maintained into later infancy through exposure to nonnative speech, also demonstrated that the decline in nonnative discriminative ability can be reversed with continued experience. That is, Pascalis et al. demonstrated that infants who were exposed to monkey faces during the 3 month period between 6 and 9 months of age continued to exhibit the ability to discriminate monkey faces at 9 months of age.

Another line of evidence that provides further proof of the contribution of developmental narrowing processes comes from studies of the "other race effect" (ORE). The ORE reflects the fact that adults find it more difficult to distinguish between the faces of people from other races than the faces of people from their own race. This finding has been interpreted as reflecting experience-dependent effects of greater exposure to individuals from one's own race than to individuals from a different race. Developmental studies have shown that the roots of the ORE can be found quite early in infancy. For example, Sangrigoli and de Schonen (2004b) tested 3-month-old Caucasian infants' preference for Caucasian and Asiatic faces following habituation to a single face from each race. They found that when infants were habituated to a Caucasian face they successfully recognized a different Caucasian face but that when they were habituated to an Asian face they did not recognize a novel Asian face. In addition, Sangrigoli and de Schonen (2004a) obtained evidence consistent with an experience-dependent interpretation of the ORE. When they gave Caucasian infants more extensive experience with other-race faces by familiarizing them with three different faces within each race the infants successfully detected differences in both Caucasian and Asiatic faces. More recently, Kelly et al. (2005) replicated the ORE in 3-month-old infants and, in addition, provided even more direct evidence of the experience-dependent nature of the ORE by showing that the ORE is absent in newborn infants. These authors concluded that selective exposure to the faces from one's own ethnic group during the first months of life leads to developmental narrowing of visual preferences to the faces from one's own ethnic group.

2.3.3 Developmental Narrowing in Face/Voice Integration

As is evident from the findings of perceptual narrowing in the speech and face processing domains, experience plays an important role in refining processing in the auditory and visual modalities. It was also noted earlier that experience plays an important role in the development of multisensory perception in that greater experience with one's own mother's face and voice increases infants' ability to integrate the two. Could it be that experience can also have the opposite effect on multisensory integration? Could it be that the lack of particular experiences can lead to a decline of an initially broadly tuned sensitivity to face-voice relations? The perceptual narrowing effects found in the speech and face processing domains show clearly that we begin life with relatively broad sensitivity in each modality and suggest that perceptual narrowing in face-voice integration may take place as well. If that is the case then this would suggest that perceptual narrowing is a domain-general developmental process.

Recently, we investigated the possibility that perceptual narrowing also plays a role in the development of multisensory perception (Lewkowicz and Ghazanfar 2006). We did so by testing the hypothesis that the range of salient multisensory relations should gradually narrow in early development and that, as a result, infants' ability to integrate nonnative faces and vocalization should be initially present early in infancy and then decline later in infancy. We put this hypothesis to empirical test by using a multisensory matching technique and stimuli from another primate species: the rhesus monkey. We presented to human infants pairs of movies in which the same macaque monkey's face could be seen on side-by-side computer monitors mouthing a "coo" call on one monitor and a "grunt" call on the other monitor. During the first two trials, infants watched these movies in silence (the side of call presentation was switched on the second trial). During the next two trials we presented the movies again but this time accompanied by either the coo or the grunt vocalization. The coo call is longer than the grunt and, thus, given that the onset of the vocalization was synchronized with the onset of both visible calls, infants could make multisensory matches on the basis of synchrony and/or duration cues. To determine when during the first year of life the decline in multisensory integration of nonnative faces and vocalizations might occur, we tested groups of 4, 6, 8, and 10 months of age. We found that 4- and 6-month old infants exhibited significantly greater looking at the matching visible call in the presence of the matching call than in its absence but that the 8- and 10-month-old infants did not. These findings are fully consistent with the previous findings of a developmental decline in responsiveness to nonnative faces and speech and show for the first time that perceptual narrowing of multisensory integration also occurs during early development and suggests that narrowing is a domain-general process.

3 The Evolution of Multisensory Vocal Perception

In this section, we broaden our discussion of multisensory social communication by asking how the kinds of multisensory perceptual mechanisms that emerge early in life during ontogenetic time might have evolved during phylogeny. We do so by considering the evolution of responsiveness to faces and voices, the primary social signals. Undoubtedly, such signals are necessary for the fast, flexible linguistic communication exhibited by humans in all societies. In order to understand how and why human vocal communication evolved and diverged from that of other primates, it is necessary to use comparative methods (Fitch 2000; Ghazanfar and Hauser 1999; Lieberman et al. 1992; Owren 2003). By comparing the vocal behavior of extant primates with human speech, one can deduce the behavioral capacities of extinct common ancestors, allowing the identification of homologies and providing clues as to the adaptive functions of such behaviors. In essence then, by exploring how extant primates perceive vocalizations, we can build a rigorous, testable framework for how multisensory speech might have evolved.

For both human and nonhuman primates, everyday social interactions often occur in noisy auditory environments in which the vocalizations of other conspecifics, heterospecifics, abiotic noise and physical obstructions can degrade the quality of auditory information. This ambient noise presents a serious obstacle to communication in all natural habitats of primates (Brown 2003). Consequently, the auditory perceptual system has evolved noise tolerant strategies to overcome these problems. Perhaps the most efficient of these evolved strategies is the audiovisual integration of vocal signals. Bimodal vocal signals in which facial expressions enhance auditory information can offer robust advantages in detection, discrimination and learning, as has been shown for multisensory signals in other domains and taxonomic groups (Rowe 1999). The classic example comes from human speech-reading: watching a speaker's face can enhance intelligibility and detection of auditory speech in noisy conditions (Bernstein et al. 2004; Cotton 1935; Grant and Seitz 2000; Schwartz et al. 2004; Sumby and Pollack 1954) and even in ideal listening conditions (Reisberg et al. 1987).

Many nonhuman primate species have large and diverse repertoires of vocalizations and facial expressions (Andrew 1962; Van Hooff 1962), and often these communication signals are co-occurring (Hauser et al. 1993; Partan 2002). The visual and auditory behavior of rhesus monkeys (*Macaca mulatta*), in particular, have been particularly well-studied (Hauser and Marler 1993; Hauser et al. 1993; Hinde and Rowell 1962; Partan 2002; Rowell and Hinde 1962). As in human speech, when rhesus monkeys produce a particular vocalization, it is often associated with a unique facial posture (Hauser et al. 1993; Partan 2002). There are three possible roles of vision in noisy auditory conditions. Seeing *who* a talker is, and then extracting indexical cues, restricts the set of possible acoustic patterns that can constitute their voice. Seeing *where* a talker is may help in binaural hearing. Seeing *when* a talker speaks can tell the observer which changes in acoustical intensity are part of the signal and which are part of the noise, since the opening of the mouth is related to the overall amplitude envelope of the speech signal. Given that vision can serve to disambiguate the auditory signal, this raises the obvious question of whether nonhuman primates can exploit the correspondence between facial postures and vocal output?

3.1 Facial Movement and Vocal Acoustics Are Linked

In primates (including humans), vocalizations are produced by coordinated movements of the lungs, larynx (vocal folds), and the supralaryngeal vocal tract (Fitch and Hauser 1995). The vocal tract consists of the column of air derived from the pharynx, mouth and nasal cavity. The source signals (sounds generated by the lungs and larynx) travel through the vocal tract and are *filtered* according to its shape, resulting in *vocal tract resonances* or *formants* discernable in the spectra of some vocalizations (for nonhuman primates, see: Fitch 1997; Owren et al. 1997; Rendall et al. 1998).

Investigators have demonstrated that human vocal tract motion necessarily results in the predictable deformation of the face around the oral aperture and other parts of the face (Jiang et al. 2002; Yehia et al. 1998; 2002). In fact, the spectral envelope of a speech signal can be predicted by the 3-dimensional motion of the face alone (Yehia et al. 1998), as can the motion of the tongue (an articulator that is not necessarily coupled with the face) (Jiang et al. 2002; Yehia et al. 1998). The spatiotemporal behavior of the vocal tract articulators involved in sound production constrains the shape and time-course of visible orofacial behavior. Such speech-related facial motion, distributed around and beyond the mouth, is what is used in the bimodal integration of audiovisual speech signals by humans. For example, human adults automatically link high-pitched sounds to facial postures producing an /i/ sound and low-pitched sounds to faces producing an /a/ sound (Kuhl et al. 1991). Humans, moreover, can identify vocal sounds when the mouth is masked or without directly looking at the mouth presumably by using such facial motion cues (Preminger et al. 1998).

In nonhuman primate vocal production, a similar link between acoustic output and facial dynamics is evident. Different rhesus monkey vocalizations are produced with unique lip configurations and mandibular positions and the motion of such articulators influences the acoustics of the signal (Hauser and Ybarra 1994; Hauser et al. 1993). Coo calls, like /u/ in speech, are produced with the lips protruded, while screams, like the /i/ in speech, are produced with the lips retracted. The jaw position and lip configuration affect the formant frequencies independent of the source frequency (Hauser and Ybarra 1994; Hauser et al. 1993). Moreover, as in humans, the articulation of these expressions has visible consequences on facial motion beyond the oral region. Grimaces, produced during scream vocalizations for instance, cause the skin-folds around the eyes to increase in number. In addition to these production-related facial movements, some vocalizations are associated with visual cues that are not directly related to the articulatory movement. Threat vocalizations, for instance, are produced with intense staring, eyebrows raised and ears often pulled back (Partan 2002). Head position and motion (e.g., chin up versus chin down versus neutral position) also varies according to vocal expression type (Partan 2002). Thus, it is likely that many of the visual motion cues that humans use for speech-reading are present in macaque monkeys as well.

3.2 Matching Faces to Voices

As indicated earlier, the ability to perceive multisensory communicative signals emerges early in human development. To reiterate, beginning in infancy humans are able to perceive various types of audio-visual relations, including face-voice relations. As indicated earlier, the development and ultimate form of this ability is shaped by progressive as well as regressive developmental processes. The former lead to the acquisition of increasingly more complex and sophisticated multisensory perceptual abilities whereas the latter lead to a decline in initially broadly tuned multisensory integration abilities that turn out to be unnecessary in the organism's natural ecology. Empirical evidence also has shown that non-human primates rely on multisensory signals for social communication (Partan and Marler 1999; Rowe 1999). Until recently, however, evidence that non-human primates can perceive the correspondence between the visual and auditory components of vocal signals was not available. To determine whether they non-human primates can perceive face-voice relations, we used the preferential-looking method that has previously been used successfully to test for crossmodal perception in pre-linguistic infants (Kuhl and Meltzoff 1984; Patterson and Werker 2003; Spelke 1976) and thus it is perfect for testing the natural capacities of non-human primates without training or reward. In essence, the method involves presenting two side-by-side digital videos of the same conspecific individual (only the head of the stimulus animal was visible) articulating two different calls to subjects seated in front of two LCD monitors. The sound track, corresponding to one of the two facial postures, was played through a hidden loudspeaker placed just behind and between the two monitors. Each video was 2-seconds in duration and the two videos were synchronized to the onset of the vocal sound. A trial consisted of the two videos played in a continuous loop for 1-minute with one of the two vocalizations also played in a loop through the speaker. The trial began when the subject fixated his gaze centrally (i.e., when the subject's eyes were directed towards the camera). The dependent measures were percentage of total looking time to the match video.

We tested whether rhesus monkeys could recognize the auditory-visual correspondences between their "coo" and "threat" calls (Ghazanfar and Logothetis 2003). These are among the most frequently produced calls in the rhesus monkey repertoire, both in the wild and in captivity. Coo calls are tonal signals of relatively long duration and are produced in many affiliative contexts, including group movements, separation and feeding; threat calls, in contrast, are noisy, short-duration pulsatile calls produced during agonistic encounters (Hauser and Marler 1993; Rowell and Hinde 1962). Each of these calls is associated with a unique facial posture. We hypothesized that the vocalization heard from the central speaker would systematically influence the duration of subjects' visual fixations on the two screens. Specifically, we predicted that if rhesus monkeys recognized the correspondence between the heard sound and the appropriate facial posture, then, overall, they would spend more time looking at the match video. The results supported our prediction. The mean percentage of looking time devoted to the match was 67.3%; this was significantly greater than the 50% chance level. Furthermore, all eleven subjects looked longer at the match face. When comparing looking durations between coos and threats, subjects did look longer when the match face was articulating the coo call than when the match face was the threat call; however, separate analyses revealed that looking preferences were still significant for coo calls alone and threat calls alone.

Although these data demonstrated that monkeys can match faces to voices, in this particular experiment they may have done so using very low-level stimulus cues. The coo and threat calls differ dramatically in their duration—the threat is almost invariably shorter than the coo. Thus, it is possible that the monkeys may have used an *amodal* cue—matching the duration of mouth movements to the duration of the auditory component (Lewkowicz 2002). This makes matching easy because the particular facial expression (i.e., the morphology of the mouth) is incidental and not necessary for making the match. Two other studies from different research groups eliminate this confound. In one study, a chimpanzee was trained to do an auditory-visual matching-to-sample task in which she was required to choose the vocalizer from two test movies in response to hearing a chimpanzee's vocalization (Izumi and Kojima 2004). This tested for individual recognition and since the method temporally separates the visual and auditory cues, the chimpanzee cannot use duration or other simple cues to do the task. The chimpanzee could match the heard vocalization with the correct face regardless of whether the face was silent or vocalizing. In another experiment with the same chimpanzee, two different types of vocalizations from the same individual were used as stimuli (Izumi and Kojima 2004). The chimpanzee again recognized that certain heard vocalizations belong to particular facial gestures. The second study used the preferential looking method to test capuchin monkeys' ability to match not only faces and voices from their own species but also from macaques and humans (Evans et al. 2005). Though the subjects matched voices to the correct faces overall, the data are inconclusive with regard to whether this effect was driven by stimuli from the same species and/or the other two species. In other words, it is still unclear whether capuchins can match heterospecific faces to voices.

3.2.1 Matching Number of Voices to Faces

As alluded to above, one context in which it is enormously helpful to combine auditory and visual communication signals is during situations in which there are multiple, simultaneous speakers. In a follow up study, we asked whether rhesus monkeys could segregate competing voices of equal duration and then match the number of voices they heard with the number of faces seen on one screen or the other. This would be akin to the way humans use audiovisual information in a cocktail party. We again employed a preferential looking technique to test whether monkeys would preferentially attend to dynamic visual displays featuring the number of unfamiliar conspecifics they simultaneously heard vocalizing. We chose to test discrimination between the quantities two versus three because these were the quantities used in all previous studies of this sort with human infants. Each of 20 monkey subjects was seated in front of two LCD monitors and a hidden speaker located between the monitors. One monitor displayed a dynamic 1-second video of 2 simultaneously vocalizing monkey faces, while the other monitor displayed a dynamic 1-second video of 3 simultaneously vocalizing monkey faces. Each video played in a continuous loop for 60 seconds. The two videos contained two common animals such that the 2-animal display was a subset of the 3-animal display. Videos were edited so that the onset and offset of all individuals' mouth movements were synchronous. Synchronous with the videos, subjects heard either 2 or 3 of these monkeys simultaneously producing coo calls. Individual coos were equated for duration and composite auditory stimuli were equated for amplitude. Since all visual and auditory components were identical in duration and synchronized, the subjects could not use amodal cues to make a match. Thus, our study asked whether monkeys would spontaneously preferentially attend to a visual stimulus that was numerically equivalent to the number of coo calls they heard.

Monkeys spent a greater proportion of time looking at the display that numerically matched the number of vocalizers they heard compared to the numerically nonmatching display. Monkeys looked at the matching display for 60% of the total time that they spent looking at either screen, which differed significantly from chance. The effect held for the monkeys who heard two calls and for the monkeys who heard three calls. Fifteen out of the 20 monkeys tested looked longer at the matching display than at the nonmatching display. These results suggest that rhesus monkeys segregated two or three simultaneously presented vocalizations and detected the numerical correspondences between the calls they heard and the vocalizing faces they saw. In other words our results showed that, rhesus monkeys can perceive the equivalence between the number of voices they hear and the number of faces they see without any explicit training and that they are capable of concurrent stream segregation of voices with overlapping spectra at a level comparable to that of humans. This kind of spontaneous, multisensory number representation in non-human animals is quite impressive and corresponds to similar adult human nonverbal number representations (Barth et al. 2003). Also impressive is the finding that rhesus monkeys can segregate simultaneously presented conspecific coo vocalizations despite the fact that the power spectra of the calls are highly overlapping Indeed, this ability is on par with humans' capacity to perceptually separate voices on the basis of pitch differences (i.e., fundamental frequency) and harmonicity (Brokx and Nooteboom 1982; Summerfield et al. 1992). This is notable because a previous study found that highly trained monkeys could discriminate concurrent sequences of *artificial* sounds only when their frequency ranges did not overlap (Izumi 2002). When the latter findings are considered together with our findings they suggest that segregation of conspecific coo vocalizatiosn is easier probably because of their far greater ecological salience as well as because of the far greater experience that animals have with such acoustic inputs during their normal development.

3.2.2 Detecting Body Size via Vocal Tract Resonances

Vowels and consonants, the essential phonetic elements of all human speech, differ in their resonances, or *formants*, produced by the vocal tract—the nasal and oral cavities above the vocal folds. During vocal production in humans and other primates, pulses of air generated by the rapid movement of the vocal folds produce an acoustic signal. As the signal passes through the vocal tract, they excite resonances in the vocal tract resulting in the enhancement of particular frequency bands; these are the formants. The ability to create and perceive a wide variety of different formant patterns is a prerequisite for human speech. This fact raises the following two questions. First, how did the ability to perceive differences between formant patterns arise? Second, which role, if any, did formants play in prelinguistic primates? The answer to these questions may lie not so much in the fact that formant patterns are important phonetic elements of speech that allow us to distinguish different vowel sounds but that they also carry important information related to physical characteristics of the individual speaker.

Previous behavioral studies demonstrated that trained baboons (Heinz and Brady 1988) and macaques (Le Prell et al. 2001; Sinnott 1989; Sommers et al. 1992) can discriminate different human vowel sounds presumably based on formant frequency differences. Recently, Fitch and Fritz (2006) have significantly extended these findings by showing that rhesus monkeys can, without training, discriminate differences in the formant structure of their own conspecific calls. However, a demonstration that particular sorts of features appear in species-typical vocalizations or that animals can attend to such features (though of great importance), is not equivalent to showing them to be functionally significant to the animals in question. The functional significance of formants in monkey vocalizations was first suggested by the study of Owren (1990a; 1990b) who showed that trained vervet monkeys could use formants to distinguish between their alarm calls (akin to the way in which humans may discriminate speech sounds).

The potential role of formants as indexical cues in rhesus monkey vocalizations was tested in a recent study by our group (Ghazanfar et al. 2007). We again used a preferential looking method to test the hypothesis that rhesus monkeys use formants as acoustic cues to assess age-related body size differences among conspecifics. The monkeys were presented with two videos. The videos showed an older, large and a younger, small monkey producing a "coo" call, but with the original sound-track removed. A synthetic audio track, originating from a third monkey, was manipulated so as to simulate either a large or a small monkey based only on changing the position of the formants. This audio track was then played simultaneously with the videos. Surprisingly, monkeys spent the most time looking at the video with the monkey that matched in size what they heard in the audio track.

The results of our experiments suggest that rhesus monkeys can not only spontaneously discriminate changes in formant structure within a call type (a la (Fitch and Fritz 2006)), but can also use these differences in formant structure as indexical cues—to assess the age-related size of a conspecific individual. Although body size is just one indexical cue among many that may be encoded in the formant frequencies of monkeys, our data show that, as in humans (Smith and Patterson 2005; Smith et al. 2005), acoustic cues that are the product of vocal tract length can be used to estimate body size. These data are the first direct evidence for the hypothesis that formants embedded in the acoustic structure of nonhuman primate calls provide cues to the physical characteristics of the vocalizer (Fitch 1997; Owren et al. 1997; Rendall et al. 1998). Our data further suggest that the use of formant cues in the perception of vowel sounds by humans in a linguistic context emerged gradually, perhaps for other functional reasons, over the course of human evolution. Perception of indexical cues, such as age-related body size, via formants in vocalizations may be one functional link between the vocalizations of human and nonhuman primates.

3.3 Eye Movements While Viewing Vocalizing Conspecifics

Watching a speaker's face can enhance perception of auditory speech under ideal (Reisberg et al. 1987) and compromised (Cotton 1935); (Sumby and Pollack 1954) listening conditions, raising the question of what cues are being used in visual speech perception. One method for investigating the behavioral strategies involved in facial-vocal processing is the measurement of eye movement patterns. Recently, studies of human subjects have examined observers' eye movements while viewing talkers in a naturalistic setting or under different listening conditions, including varying levels of background noise (Vatikiotis-Bateson et al. 1998), competing voices (Rudmann et al. 2003), and silence (i.e., speech-reading with no audio track) (Lansing and McConkie 1999; 2003). When human subjects are given *no* task or instruction regarding what acoustic cues to attend, they will consistently look at the eye region more than the mouth when viewing videos of human speakers. However, when subjects are required to perform a specific task, then eye movement patterns are task-dependent. For example, when required to attend to speech-specific aspects of the communication signal (e.g., phonetic details in high background noise, word identification or segmental cues), humans direct significantly more fixations to the mouth region than the eye region (Lansing and McConkie 2003; Vatikiotis-Bateson et al. 1998). When, however, subjects are asked to focus on prosodic cues or to make social judgments based on what they see/hear, they direct their gaze more often toward the eyes than the mouth (Buchan et al. 2004; 2005); Lansing and McConkie 1999). The evolution of sensorimotor mechanisms that analyze and integrate facial and vocal expressions is likely an innovation that is not specific to human speech perception (Ghazanfar and Santos 2004). We don't know, however, whether humans and monkeys use the same sensorimotor processes when they view vocalizing conspecifics.

To characterize the potential similarities and differences between monkey and human audiovisual communication, we investigated the eye movement patterns of rhesus monkeys while they viewed digitized videos of conspecifics producing vocalizations (Ghazanfar et al. 2006). We generated video sequences of monkeys vocalizing and varied the listening conditions by modifying the audio track. In one experiment, we compared responses to normal movie sequences with sequences in which the audio track was either silenced or where the auditory component of the vocalizations were paired with the incorrect facial posture (i.e., mismatched). The monkey subjects were not required to perform a task, but simply free-viewed the videos in whatever spontaneous manner they chose. Under all listening conditions, our monkey subjects spent most of their time inspecting the eye region relative to the mouth. When they did fixate on the mouth, it was highly correlated with the onset of mouth movements. Finally, there was no relationship between the number or duration of fixations with respect to call type. We concluded, therefore, that the auditory component has no influence on eye movement patterns of monkeys viewing vocalizing conspecifics.

Our findings have striking parallels with what we know about human eye movement patterns during speech-reading. In both species, the greater number of fixations fall in the eye region than in the mouth region when subjects are required simply to view vocalizing conspecifics, to attend to emotion-related cues or to make social judgments (Buchan et al. 2004; 2005). Even during visual speech alone (no auditory component), when subjects are asked to attend to prosodic cues, they will look at the eyes more than the mouth (Lansing and McConkie 1999). Furthermore, like human observers (Lansing and McConkie

2003), monkeys look at the eyes *before* they look at the mouth and their fixations on the mouth are tightly correlated with mouth movement. For instance, Lansing and McConkie (2003) reported that, regardless of whether it was visual or audiovisual speech, subjects asked to identify words increased their fixations onto the mouth region with the onset of facial motion. The same was true for rhesus monkeys (Ghazanfar et al. 2006).

The precise developmental emergence of the types of visual scanning patterns found in adults interacting with conspecifics is not completely clear at this point. It does appear, however, that adult-like eye movement patterns emerge during the early part of infancy. Initially, during the first few months of life, infants exhibit "sticky" attention in that they tend to stare at particular visual attributes and find it difficult to disengage from them once they fixate on them. With further development, however, this sticky attention gives way to more flexible fixation patterns and by around the fourth month of life infants begin to exhibit more flexible visual attention in that they can now shift gaze between different stimulus attributes or different stimuli (Frick et al. 1999). With specific regard to scanning of faces, it is difficult to tell at this point whether infants direct more of their gaze at the mouth region when they hear speech than when they do not. Only one early study (Haith et al. 1977) investigated infants' looking at vocalizing faces. Overall, the findings from this study indicated that infants between 3 and 11 weeks of age looked mostly at the eye region and that when a voice accompanied the face, 7- and 9- to 11-week-old infants exhibited marginally longer looking at the eyes when listening to speech. One recent study (Hunnius and Geuze 2004) investigated infant scanning patterns of silent talking faces and found that infants between 6 and 26 weeks of age exhibited increasing fixation of the mouth region rather than the eye region. Given that the vocal part of the event was absent in this study, the greater fixation of the mouth probably reflects a violation of expectancy. That is, because infants have a great deal of experience with talking faces right from birth, it is likely that they were expecting to hear the voice and when they did not hear they searched for it by looking at the region of the mouth that the voice normally comes from. In sum, it is still not clear what effects speech might have on the development of face perception and whether different types of speech and/or the specific temporal relations between the visual and auditory attributes of a talking face might have on infant scanning of talking faces.

4 Neural Mechanisms of Face/Voice Integration

Our current knowledge of bimodal integration of visual and auditory primate vocal signals in the brain is derived almost exclusively from human neuroimaging studies of audiovisual speech. The imaging data consistently show that the superior temporal sulcus and auditory cortex (including primary auditory cortex) seem to play important roles in the integration process (Calvert 2001). Animal models of multisensory integration have largely focused on the role of the superior colliculus in the spatial and temporal integration of simple artificial stimuli (see Stein 1998 for a recent review). Although many "principles" of multisensory integration have been identified for neurons in the superior colliculus, it is unknown to what extent these principles apply to complex, naturalistic stimuli in the primate neocortex. Human neuroimaging data of audiovisual speech have added much to our knowledge of its underlying cortical circuitry and have generated several interesting hypotheses; however, the nature of imaging signals precludes a direct understanding of how this network functions and interacts. Problems inherent in these imaging and event-related potential studies include the inability to distinguish between inhibitory versus excitatory responses, poor source localization and/or poor temporal resolution.

To bridge these epistemic gaps, recent studies have investigated dynamic face/voice integration in the superior temporal sulcus (Barraclough et al. 2005), the prefrontal cortex (Sugihara et al. 2006), and presumptive unimodal auditory areas (Ghazanfar et al. 2005) of the rhesus monkey neocortex using the species' natural communication signals. Unlike pairings of artificial stimuli, audiovisual vocalizations are ethologically-relevant and thus may tap into specialized neural mechanisms (Ghazanfar and Santos 2004) or, minimally, integrative mechanisms for socially-learned audio-visual associations. In addition, the spatiotemporal complexity of facial and vocal signals may uncover principles of multisensory integration that cannot be revealed with the use of simple, static stimuli.

The superior temporal sulcus has long been known to have neurons that are responsive to visual, auditory and/or somatosensory stimuli (Benevento et al. 1977; Bruce et al. 1981; Hikosaka et al. 1988; Schroeder and Foxe 2002). Recently, Barraclough et al. (2005) systematically investigated the integrative properties of single neurons in STS using biologically-relevant dynamic stimuli such as ripping paper, chewing and vocalizations. They found that 23% of neurons responsive to the sight of biological motion could be significantly modulated by the corresponding auditory component either in the form of enhancement or suppression. Recently, Sugihara et al. (2006) have reported that neurons in the ventrolateral prefrontal cortex also integrate face/voice stimuli also in the form of both enhancement and suppression. Approximately 46% of neurons that were visually or auditory responsive were multisensory.

Surprisingly, auditory cortex has been shown to be polysensory as well, responding to somatosensory as well as visual stimuli (Fu et al. 2003; Kayser et al. 2005; Schroeder and Foxe 2002). We investigated the *integrative* properties of the auditory cortex by having monkey subjects view unimodal and bimodal versions of two different species-typical vocalizations ("coos" and "grunts") while we recorded the mean extracellular field potential (i.e., unit and sub-threshold neural activity) from the core and lateral belt (LB) region of auditory cortex (Ghazanfar et al. 2005). Our data unequivocally demonstrated that LFPs in the auditory cortex are capable of multisensory integration of facial and vocal signals in monkeys (Ghazanfar et al. 2005). The vast majority of responses were specific to face + voice integration and such integration could take the form of either enhancement or suppression, although enhanced responses were more common. These enhanced responses were biased towards one of the two call types used in our study, the “grunt”. We suspect that this bias towards grunt calls is related to the fact the grunts (relative to coos) are often produced during intimate, one-to-one social interactions.

5 Conclusions

Using a systems perspective, the current chapter discussed multisensory perception of communicative signals by considering the proximal behavioral and neural mechanisms underlying multisensory perception and integration, its ontogeny, and its phylogeny. We have shown that multisensory perception and integration develops early in life, that it is a pervasive feature of primate behavioral and neural organization, and that it has evolutionary roots and value. Specifically, we showed that multisensory perception emerges during infancy, that it consists of the gradual development of increasingly more sophisticated integration abilities, and that experience plays a key role in the development of these abilities. We have also shown that these abilities are evolutionarily old in that nonhuman adult primates also exhibit them and discussed some of the recent evidence on the neural mechanisms underlying these abilities. This evidence indicates that the neural mechanisms underlying audiovisual integration of biologically-relevant signals reside in homologous networks in humans and monkeys, embedded within the temporal, parietal and frontal lobes. Although much more comparative work on other primate and non-primate species is needed, the macaque monkey appears to be an outstanding model system for the neuro-cognitive investigation of auditory-visual communication. To date, the data suggest that the last common ancestor of humans and anthropoid monkeys adopted an integrated audiovisual mode of vocal communication and future studies should further investigate the limits of this ability as well as its development.

References

Andrew RJ (1962) The origin and evolution of the calls and facial expressions of the primates. Behaviour 20:1–109

Aslin RN, Pisoni DB, Hennessy BL, Percy AJ (1981) Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. Child Development 52:1135–1145

Barraclough NE, Xiao DK, Baker CI, Oram MW, Perrett DI (2005) Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. Journal of Cognitive Neuroscience 17:377–391

Barth H, Kanwisher N, Spelke ES (2003) The construction of large number representations in adults. Cognition 86:201–221

Benevento LA, Fallon J, Davis BJ, Rezak M (1977) Auditory-visual interactions in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. Experimental Neurology 57:849–872

Bernstein LE, Auer ET, Takayanagi S (2004) Auditory speech detection in noise enhanced by lipreading. Speech Communication 44:5–18

Best CT, McRoberts GW, LaFleur R, Silver-Isenstadt J (1995) Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. Infant Behavior & Development 18:339–350

Birch HG, Lefford A (1963) Intersensory development in children. Monographs of the Society for Research in Child Development 25

Birch HG, Lefford A (1967) Visual differentiation, intersensory integration, and voluntary motor control. Monographs of the Society for Research in Child Development 32:1–87

Brokx JP, Nooteboom SG (1982) Intonation and the perceptual separation of simultaneous voices. Journal of Phonetics 10:23–36

Brown CH (2003) Ecological and physiological constraints for primate vocal communication. In: Ghazanfar AA (ed) Primate audition: Ethology and neurobiology. CRC Press, Boca Raton, FL, pp 127–150

Bruce C, Desimone R, Gross CG (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. Journal of Neurophysiology 46:369–384

Buchan JN, Pare M, Munhall KG (2004) The influence of task on gaze during audiovisual speech perception. Journal of the Acoustical Society of America 115:2607

Buchan JN, Pare M, Munhall KG (2007) Spatial statistics of gaze fixations during dynamic face processing. Social Neuroscience 2:1–13

Bushnell EW (1994) A dual-processing approach to cross-modal matching: Implications for development. In: Lewkowicz DJ, Lickliter R (eds) The development of intersensory perception: Comparative perspectives. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, pp 19–38

Calvert GA (2001) Crossmodal processing in the human brain: Insights from functional neuroimaging studies. Cerebral Cortex 11:1110–1123

Cohen LB, Strauss MS (1979) Concept acquisition in the human infant. Child Development 50:419–424

Cotton JC (1935) Normal "visual hearing". Science 82:592–593

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

DeCasper AJ, Fifer WP (1980) Of human bonding: Newborns prefer their mothers' voices. Science 208:1174–1176

Dufour V, Pascalis O, Petit O (2006) Face processing limitation to own species in primates: A comparative study in brown capuchins, Tonkean macaques and humans. Behavioural Processes 73:107–113

Eimas PD, Siqueland ER, Jusczyk P, Vigorito J (1971) Speech perception in infants. Science 171:303–306

Evans TA, Howell S, Westergaard GC (2005) Auditory-visual cross-modal perception of communicative stimuli in tufted capuchin monkeys (Cebus apella). Journal of Experimental Psychology, Animal Behavior Processes 31:399–406

Fernald A, Pinto JP, Swingley D, Weinberg A, McRoberts GW (1998) Rapid gains in speed of verbal processing by infants in the 2nd year. Psychological Science 9:228–231

Fitch WT (1997) Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. Journal of the Acoustical Society of America 102:1213–1222

Fitch WT (2000) The evolution of speech: A comparative review. Trends Cognitive Sciences 4:258–267

Fitch WT, Fritz JB (2006) Rhesus macaques spontaneously perceive formants in conspecific vocalizations. Journal of the Acoustical Society of America 120:2132–2141

Fitch WT, Hauser MD (1995) Vocal production in nonhuman-primates—Acoustics, physiology, and functional constraints on honest advertisement. American Journal of Primatology 37:191–219

Frick JE, Columbo J, Saxon TF (1999) Individual and developmental differences in disengagement of fixation in early infancy. Child Development 70:537–548

Fu KMG, Johnston TA, Shah AS, Arnold L, Smiley J, Hackett TA, Garraghty PE, Schroeder CE (2003) Auditory cortical neurons respond to somatosensory stimulation. Journal of Neuroscience 23:7510–7515

Ghazanfar AA, Hauser MD (1999) The neuroethology of primate vocal communication: Substrates for the evolution of speech. Trends in Cognitive Sciences 3:377–384

Ghazanfar AA, Logothetis NK (2003) Facial expressions linked to monkey calls. Nature 423:937–938

Ghazanfar AA, Santos LR (2004) Primate brains in the wild: The sensory bases for social interactions. Nature Reviews Neuroscience 5:603–616

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends in Cognitive Sciences 10:278–285

Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. Journal of Neuroscience 25:5004–5012

Ghazanfar AA, Nielsen K, Logothetis NK (2006) Eye movements of monkeys viewing vocalizing conspecifics. Cognition 101:515–529

Ghazanfar AA, Turesson HK, Maier JX, van Dinther R, Patterson RD, Logothetis NK (2007) Vocal tract resonances as indexical cues in rhesus monkeys. Current Biology 17:425–430

Gibson E (1969) Principles of perceptual learning and development. Appleton, New York

Gibson EJ (1984) Perceptual development from the ecological approach. In: Lamb ME, Brown AL, Rogoff B (eds) Advances in developmental psychology. Lawrence Erlbaum Associates, Hillsdale, NJ, pp 243–286

Gottlieb G (1991) Experiential canalization of behavioral development: Results. Developmental Psychology 27:35–39

Gottlieb G (1992) Individual development & evolution: The genesis of novel behavior. Oxford University Press, New York

Gottlieb G (1996) Developmental psychobiological theory. In: Cairns RB, Elder GH Jr (eds) Developmental science. Cambridge studies in social and emotional development. Cambridge University Press, New York, pp 63–77

Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. Journal of the Acoustical Society of America 108:1197–1208

Haith MM, Bergman T, Moore MJ (1977) Eye contact and face scanning in early infancy. Science 198:853–855

Hauser MD, Marler P (1993) Food-associated calls in rhesus macaques (Macaca-Mulatta). 1. Socioecological factors. Behavioral Ecology 4:194–205

Hauser MD, Ybarra MS (1994) The role of lip configuration in monkey vocalizations—Experiments using xylocaine as a nerve block. Brain and Language 46:232–244

Hauser MD, Evans CS, Marler P (1993) The role of articulation in the production of rhesus-monkey, Macaca-Mulatta, vocalizations. Animal Behaviour 45:423–433

Heinz RD, Brady JV (1988) The acquisition of vowel discriminations by nonhuman primates. Journal of the Acoustical Society of America 84:186–194

Hikosaka K, Iwai E, Saito HA, Tanaka K (1988) Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. Journal of Neurophysiology 60:1615–1637

Hinde RA, Rowell TE (1962) Communication by posture and facial expressions in the rhesus monkey (*Macaca mulatta*). Proceedings of the Zoological Society London 138:1–21

Holt EB (1931) Animal drive and the learning process, vol 1. Holt, New York

Hunnius S, Geuze RH (2004) Developmental Changes in Visual Scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. Infancy 6:231–255

Huxley TH (1863) Evidences as to man's place in nature. Williams and Norgate, London

Izumi A (2002) Auditory stream segregation in Japanese monkeys. Cognition 82: B113–B122

Izumi A, Kojima S (2004) Matching vocalizations to vocalizing faces in a chimpanzee (Pan troglodytes). Animal Cognition 7:179–184

Jiang JT, Alwan A, Keating PA, Auer ET, Bernstein LE (2002) On the relationship between face movements, tongue movements, and speech acoustics. Eurasip Journal on Applied Signal Processing 2002:1174–1188

Jusczyk PW (1997) The discovery of spoken language. MIT Press, Cambridge, MA

Kayser C, Petkov CI, Augath M, Logothetis NK (2005) Integration of touch and sound in auditory cortex. Neuron 48:373–384

Kelly DJ, Quinn PC, Slater AM, Lee K, Gibson A, Smith M, Ge L, Pascalis O (2005) Three-month-olds, but not newborns, prefer own-race faces. Developmental Science 8: F31–F36

Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. Science 218:1138–1141

Kuhl PK, Meltzoff AN (1984) The intermodal representation of speech in infants. Infant Behavior & Development 7:361–381

Kuhl PK, Williams KA, Meltzoff AN (1991) Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. Journal of Experimental Psychology, Human Perception and Performance 17:829–840

Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B (1992) Linguistic experience alters phonetic perception in infants by 6 months of age. Science 255:606–608

Kuhl PK, Tsao FM, Liu HM (2003) 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 100:9096–9101

Kuo ZY (1976) The dynamics of behavior development: An epigenetic view. Plenum, New York

Lansing CR, McConkie GW (1999) Attention to facial regions in segmental and prosodic visual speech perception tasks. Journal of Speech Language and Hearing Research 42:526–539

Lansing IR, McConkie GW (2003) Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. Perception & Psychophysics 65:536–552

Le Prell CG, Niemiec AJ, Moody DB (2001) Macaque thresholds for detecting increases in intensity: Effects of formant structure. Hearing Research 162:29–42

Lewis MD (2000) The promise of dynamic systems approaches for an integrated account of human development. Child Development 71:36–43

Lewkowicz DJ (1986) Developmental changes in infants' bisensory response to synchronous durations. Infant Behavior & Development 9:335–353

Lewkowicz DJ (1992a) Infants response to temporally based intersensory equivalence—The effect of synchronous sounds on visual preferences for moving stimuli. Infant Behavior & Development 15:297–324

Lewkowicz DJ (1992b) Infants responsiveness to the auditory and visual attributes of a sounding moving stimulus. Perception & Psychophysics 52:519–528

Lewkowicz DJ (1994) Development of intersensory perception in human infants. In: Lewkowicz DJ, Lickliter R (eds) The development of intersensory perception: Comparative perspectives. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, pp 165–203

Lewkowicz DJ (1996) Perception of auditory-visual temporal synchrony in human infants. Journal of Experimental Psychology, Human Perception and Performance 22:1094–1106

Lewkowicz DJ (2000a) The development of intersensory temporal perception: An epigenetic systems/limitations view. Psychological Bulletin 126:281–308

Lewkowicz DJ (2000b) Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. Child Development 71:1241–1257

Lewkowicz DJ (2002) Heterogeneity and heterochrony in the development of intersensory perception. Cognitive Brain Research 14:41–63

Lewkowicz DJ (2003) Learning and discrimination of audiovisual events in human infants: The hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. Developmental Psychology 39:795–804

Lewkowicz DJ, Ghazanfar AA (2006) The decline of cross-species intersensory perception in human infants. Proceedings of the National Academy of Sciences of the United States of America 103:6771–6774

Lewkowicz DJ, Turkewitz G (1980) Cross-modal equivalence in early infancy: Auditory-visual intensity matching. Developmental Psychology 16:597–607

Lickliter R (2000) An ecological approach to behavioral development: Insights from comparative psychology. Ecological Psychology 12:319–334

Lieberman P, Laitman JT, Reidenberg JS, Gannon PJ (1992) The anatomy, physiology, acoustics and perception of speech: Essential elements in analysis of the evolution of human speech. Journal of Human Evolution 23:447–467

Ludemann PM, Nelson CA (1988) Categorical representation of facial expressions by 7-month-old infants. Developmental Psychology 24:492–501

Marcus G, Vijayan S, Rao S, Vishton P (1999) Rule learning by seven-month-old infants. Science 283:77–80

Marks L (1978) The unity of the senses. Academic Press, New York

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:229–239

Morton J, Johnson MH (1991) CONSPEC and CONLERN: A two-process theory of infant face recognition. Psychological Review 98:164–181

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 & Performance 24:756–766

Neil PA, Chee-Ruiter C, Scheier C, Lewkowicz DJ, Shimojo S (2006) Development of multisensory spatial integration and perception in humans. Developmental Science 9:454–464

Nelson CA (2001) The development and neural bases of face recognition. Infant & Child Development 10:3–18

Owren MJ (1990a) Acoustic classification of alarm calls by vervet monkeys (Cercopithecus-Aethiops) and humans (Homo-Sapiens). 1. Natural calls. Journal of Comparative Psychology 104:20–28

Owren MJ (1990b) Acoustic classification of alarm calls by vervet monkeys (Cercopithecus-Aethiops) and humans (Homo-Sapiens). 2. Synthetic calls. Journal of Comparative Psychology 104:29–40

Owren MJ (2003) Vocal production and perception in nonhuman primates provide clues about early hominids and speech evolution. ATR Symposium HIS Series 1:1–19

Owren MJ, Seyfarth RM, Cheney DL (1997) The acoustic features of vowel-like grunt calls in chacma baboons (Papio cyncephalus ursinus): Implications for production processes and functions. Journal of the Acoustical Society of America 101:2951–2963

Partan S, Marler P (1999) Communication goes multimodal. Science 283:1272–1273

Partan SR (2002) Single and multichannel signal composition: Facial expressions and vocalizations of rhesus macaques (*Macaca mulatta*). Behaviour 139:993–1027

Pascalis O, Bachevalier J (1998) Face recognition in primates: A cross-species study. Behavioural Processes 43:87–96

Pascalis O, de Haan M, Nelson CA (2002) Is face processing species-specific during the first year of life? Science 296:1321–1323

Pascalis O, Scott LS, Kelly DJ, Shannon RW, Nicholson E, Coleman M, Nelson CA (2005) Plasticity of face processing in infancy. Proceedings of the National Academy of Sciences of the United States of America 102:5297–5300

Patterson ML, Werker JF (2002) Infants' ability to match dynamic phonetic and gender, information in the face and voice. Journal of Experimental Child Psychology 81:93–115

Patterson ML, Werker JF (2003) Two-month-old infants match phonetic information in lips and voice. Developmental Science 6:191–196

Piaget J (1952) The origins of intelligence in children. International Universities Press, New York

Preminger JE, Lin H-B, Payen M, Levitt H (1998) Selective visual masking in speechreading. Journal of Speech, Language, and Hearing Research 41:564–575

Reardon P, Bushnell EW (1988) Infants' sensitivity to arbitrary pairings of color and taste. Infant Behavior and Development 11, 245–250

Reisberg D, McLean J, Goldfield A (1987) Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In: Dodd B, Campbell R (eds) Hearing by eye: The psychology of lip-reading. Erlbaum, Hillsdale, NJ, pp 97–113

Rendall D, Owren MJ, Rodman PS (1998) The role of vocal tract filtering in identity cueing in rhesus monkey (*Macaca mulatta*) vocalizations. Journal of the Acoustical Society of America 103:602–614

Rowe C (1999) Receiver psychology and the evolution of multicomponent signals. Animal Behaviour 58:921–931

Rowell TE, Hinde RA (1962) Vocal communication by the rhesus monkey (*Macaca mulatta*). Proceedings of the Zoological Society London 138:279–294

Rudmann DS, McCarley JS, Kramer AF (2003) Bimodal displays improve speech comprehension in environments with multiple speakers. Human Factors 45:329–336

Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. Science 274:1926–1928

Sangrigoli S, de Schonen S (2004a) Effect of visual experience on face processing: A developmental study of inversion and non-native effects. Developmental Science 7:74–87

Sangrigoli S, de Schonen S (2004b) Recognition of own-race and other-race faces by three-month-old infants. Journal of Child Psychology and Psychiatry 45:1219–1227

Schneirla TC (1949) Levels in the psychological capacities of animals. In: Sellars RW, McGill VJ, Farber M (eds) Philosophy for the future. Macmillan, New York, pp 243–286

Schroeder CE, Foxe JJ (2002) The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. Cognitive Brain Research 14:187–198

Schwartz J-L, Berthommier F, Savariaux C (2004) Seeing to hear better: Evidence for early audio-visual interactions in speech identification. Cognition 93:B69–B78

Sinnott JM (1989) Detection and discrimination of synthetic English vowels by old-world monkeys (Cercopithecus, Macaca) and humans. Journal of the Acoustical Society of America 86:557–565

Smith DRR, Patterson RD (2005) The interaction of glottal-pulse rate and vocal-tract length in judgement of speaker size, sex and age. Journal of the Acoustical Society of America 118:3177–3186

Smith DRR, Patterson RD, Turner R, Kawahara H, Irino T (2005) The processing and perception of size information in speech sounds. Journal of the Acoustical Society of America 117:305–318

Sommers MS, Moody DB, Prosen CA, Stebbins WC (1992) Formant frequency discrimination by Japanese macaques (Macaca-Fuscata). Journal of the Acoustical Society of America 91:3499–3510

Spelke ES (1976) Infants' intermodal perception of events. Cognitive Psychology 8:553–560

Stein BE (1998) Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. Experimental Brain Research 123:124–135

Stein BE, Meredith MA (1993) The merging of the senses. MIT Press, Cambridge, MA

Streeter LA (1976) Language perception of 2-mo-old infants shows effects of both innate mechanisms and experience. Nature 259:39–41

Sugihara T, Diltz M, Averbeck BB, Romanski LM (2006) Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. Journal of Neuroscience 26:11138–11147

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26:212–215

Summerfield Q, Culling JF, Fourcin AJ (1992) Auditory segregation of competing voices: Absence of effects of FM or AM coherence. Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences 336:357–366

Thelen E, Smith LB (1994) A dynamic systems approach to the development of cognition and action. MIT Press, Cambridge, MA

Van Hooff JARAM (1962) Facial expressions of higher primates. Symposium of the Zoological Society, London 8:97–125

Vatikiotis-Bateson E, Eigsti IM, Yano S, Munhall KG (1998) Eye movement of perceivers during audiovisual speech perception. Perception & Psychophysics 60:926–940

Walker-Andrews AS (1986) Intermodal perception of expressive behaviors: Relation of eye and voice? Developmental Psychology 22:373–377

Walker-Andrews AS (1997) Infants' perception of expressive behaviors: Differentiation of multimodal information. Psychological Bulletin 121:437–456

Walker-Andrews AS, & Dickson LR (1997) Infants' understanding of affect. In: Hala S (ed) The development of social cognition. Studies in developmental psychology. Psychology Press/Erlbaum (UK) Taylor & Francis, Hove, England, pp 161–186

Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA (2004) Unifying multisensory signals across time and space. Experimental Brain Research 158:252–258

Werker JF, Tees RC (1984) Cross-language speech-perception—Evidence for perceptual reorganization during the 1st year of life. Infant Behavior & Development 7:49–63

Werner H (1973) Comparative psychology of mental development. International Universities Press, New York

Yehia H, Rubin P, Vatikiotis-Bateson E (1998) Quantitative association of vocal-tract and facial behavior. Speech Communication 26:23–43

Yehia HC, Kuratate T, Vatikiotis-Bateson E (2002) Linking facial animation, head motion and speech acoustics. Journal of Phonetics 30:555–568

8
Understanding the Dynamics of Primate Vocalization and Its Implications for the Evolution of Human Speech

Takeshi Nishimura

1 Introduction

The origin of language remains one of the most enigmatic issues for studies on human evolution and is a challenge that attracts many scholars. Any discussion of this issue has been taboo in traditional linguistics, but in the past ten years it has been examined productively through informatics, biology and cognitive science, as well as by using theoretical linguistics. The interested reader can consult the books edited by Wray (2002) and Christiansen and Kirby (2003) for details. Morphologists and paleoanthropologists have continued to debate this issue long before the successes of other disciplines. However, they have faced a great obstacle in their efforts: language *per se* cannot fossilize and leaves no archaeological traces. This is a great distinction in any starting point for paleontological studies on the origin of language and the other issues. For example, the evolution of habitual bipedal walking is accessible through examination of the cranial base, pelvis, leg or foot bones of fossil forms. Nevertheless, such difficulties have never let the scholars abandon their ambitions to challenge the enigmas surrounding the origin and evolution of language, which has doubtless contributed to the unfolding of humanity and its civilizations.

No human groups lack verbal communication by speech, whereby concepts encoded by language in one brain can be represented in another brain, although some lack other forms of communication, such as writing. Humans have a faculty to produce distinct phonemes—including vowels and consonants—sequentially and rapidly even in a short single exhalation, without any significant effort. This sophisticated feature of speech allows humans to convert much of the information encoded by language into a meaningful series of sounds with great efficiency (Lieberman 1984). The acoustic feature of speech is achieved by sophisticated manipulation of the vocal apparatuses, including the lung, larynx, tongue, jaws and lips (Lieberman and Blumstein 1988). Although humans make use of the same peripheral machinery as other mammals for vocalization, human speech

Primate Research Institute, Kyoto University, Inuyama, Aichi 484-8506, Japan

entails several important and unique modifications to anatomy *per se* and to the neurological foundations that regulate the machinery. The features affected by such modifications are available through comparative studies for extant primates, including humans. Thus, although speech *per se* is merely a kind of vocalization and is not the same as language, paleoanthropologists and morphologists have made great efforts to use these clues to evaluate the likely speech faculties of fossil hominins, and have shed light on the evolution of the language facility with which we are now endowed.

The unique anatomy and physiology of human speech has now been reevaluated using new disciplinary approaches. The vocal apparatus is composed mainly of soft tissues including a cartilaginous skeleton, muscles and ligaments, and the region of interest is almost entirely veiled by the hard tissue of the cranium and mandible (Williams 1995; Zemlin 1988). In addition, vocalizations are performed using continual and varied actions of the vocal apparatuses (Fant 1960; Lieberman and Blumstein 1988; Stevens 1998). Such anatomies and dynamic activities *per se* have prevented detailed examination of the anatomical developments and vocal physiology in living nonhuman primates. However, newer medical imaging techniques and updated knowledge of the bioacoustics are breaking down the technical limitations and have enabled us to review the established knowledge. Computed tomography (CT) and magnetic resonance imaging (MRI) have allowed physical anthropologists and primatologists to evaluate the developmental changes in the vocal apparatus in intact living subjects (Flügel and Rohen 1991; Nishimura et al. 2003; 2006). Recent evaluations of vocal physiology have shown that nonhuman primates also perform a wide range of formant patterns and formant transitions that had been believed to be unique to human speech (e.g., Riede and Zuberbühler 2003). These have suggested that such an acoustical performance requires more of a dynamic action of the vocal apparatus than expected (Fitch 2000b; Riede et al. 2006). Although in hindsight some of the anatomical and physiological uniqueness associated with human speech might have been overemphasized, these new disciplines will redefine the features of human speech relative to primate vocalizations.

This chapter briefly reviews speech physiology and anatomy. It relates the efforts made by physical anthropologists over the past forty years to reconstruct the origin of speech. Finally, it surveys recent advances in anatomical and physiological studies concerning vocal faculties in nonhuman primates. These promise further understanding of the origin and evolution of human speech and language.

2 Human Speech Physiology

Humans utter a series of distinct sounds containing voiced vowels and voiced and unvoiced consonants. The terms *voiced* and *unvoiced* mean that the sounds are produced accompanied or unaccompanied with vibrations of the vocal folds of the glottis, respectively. The vowels are quantally voiced sounds produced

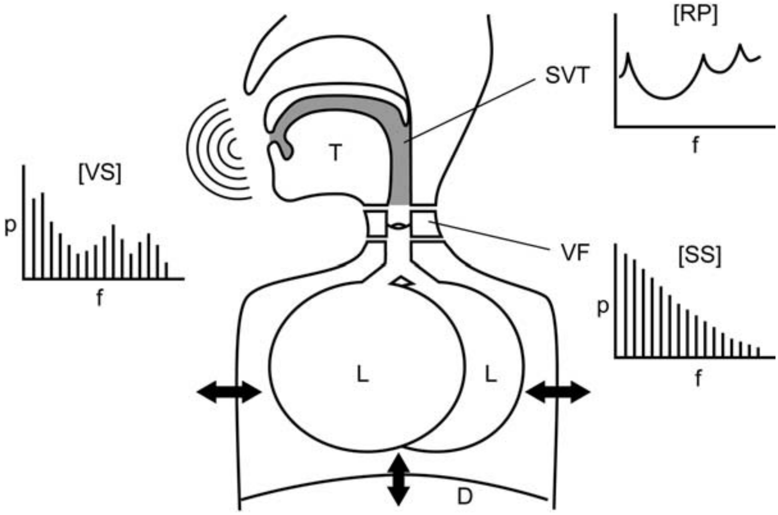


Fig. 1. Diagram for the acoustic theory in vocalizations. The sound power is the exhaled airflow produced by the compression of pulmonary volume (*arrows*). The sound sources (*SS*) are produced by the vibrations of vocal folds (*VF*) in the larynx. The supralaryngeal vocal tract (*SVT*, colored in grey) functions as the resonator for the sources to generate the quantal vowel sounds (*VS*) uttered from the mouth. The resonant properties (*RP*) of the SVT depend on topology of the tract which is modified by the tongue (*T*). *L*, lung; *D*, diaphragm.

through the resonance of the sound sources by the supralaryngeal vocal tract (SVT), and consonants are noises often followed by the vowels in speech. Humans coordinate varied movements of the vocal apparatuses in a sophisticated manner to produce distinct speech sounds sequentially and rapidly. Understanding of the complicated physiology involved advanced initially through the advancement of telephone technology (Raphael et al. 2007). These insights provide the basis for our better understanding of bioacoustics in nonhuman vocalizations. Although the consonants play a critical role in speech, the physiology explained here deal with the production of the vowels that form the foundation of speech. This is interpreted with the Source–Filter Theory (Fant 1960), which explains that the vowels are produced via three steps: exhalation of air from the lungs, phonation in the larynx and articulation in the SVT (Fig. 1: Chiba and Kajiyama 1942; Fant 1960; Stevens 1998; Titze 1994). This section provides a brief review of the physiology and associated anatomical foundations of vowel production.

2.1 Voluntary Regulation of Respiration

The sound power is the exhaled airflow produced by the compression of pulmonary volume. It is achieved by the contraction or relaxation of associated muscles in the thorax and abdomen. The transversus abdominis, rectus abdominis and pectoralis major muscles are contracted to compress the pulmonary volume to induce exhalation (Williams 1995; Zemlin 1988). In contrast, the internal and external intercostal muscles and the diaphragm are contracted to increase the pulmonary volume and induce inspiration by reducing intrathoracic relative pressure (Williams 1995; Zemlin 1988). Contraction and relaxation of these muscle sets are regulated involuntarily and automatically in tidal respiration: the exhaled airflow usually has less force toward the end of an exhalation, and exhalation and inspiration are repeated alternately with almost identical durations. However, humans can voluntarily and flexibly regulate such a counterbalance between the activities of these muscle groups to maintain a constant air pressure to the end of a single exhalation (MacLarnon and Hewitt 2004; Titze 1994). Moreover, a short inspiration is taken quickly with precise timing at the ends of sentences in speech, in contrast to tidal respiration. Thus, such a sophisticated voluntary regulation of the respiratory apparatus is a requisite to the steady production of any series of speech sounds in humans.

2.2 Phonation in the Larynx

The exhaled airflow reaches the glottis, composed of bilateral vocal folds, to run up through the narrow channels of the glottis. It usually induces symmetrical and cyclical vibrations of the vocal folds to generate a series of air puffs expelled to the SVT (Lieberman and Blumstein 1988; Titze 1994). These form sound sources, called *laryngeal sounds* or *glottal sources*, and this process is known as *phonation*. The elasticity, lengths and thicknesses of the vocal folds and the air pressure of the sub- and supralaryngeal cavities determine the quality of the laryngeal sounds, affecting the intensity, loudness and pitch of the speech sounds (Titze 1994). These physical processes are complicated and some remain under debate. Nevertheless, it is established that the physical property of the vocal folds is mostly regulated by the spatial relationship of the laryngeal skeleton with the musculature of the vocal folds. The laryngeal skeleton is composed of the thyroid, cricoid, two arytenoid and other small cartilages, but contains no bony elements. The thyroarytenoid and vocal muscles and vocal ligaments composing the vocal folds originate from beneath the suprathyroid notch of the thyroid, the so-called *Adam's apple*, and insert onto the vocal processes of the bilateral arytenoids, which cradle on the superoposterior edge of the cricoid ring. The spatial relationships of the cartilages are changeable by the contraction and relaxation of the external and internal laryngeal musculature. For example, the thyroarytenoid muscle is contracted to make the arytenoid cartilages rock into the cricoid ring to approximate the bilateral vocal folds, and to stiffen and shorten the folds, while the posterior cricoarytenoid muscle is contracted to induce contrary actions (Williams 1995; Zemlin 1988). Thus, humans voluntarily regulate the activities of the laryngeal musculature to produce laryngeal sounds appropriate for the production of sequential vowels and consonants, accompanied by the on–off regulation of such mechanisms.

2.3 Articulation in the Supralaryngeal Vocal Tract

The SVT, from the glottis to the lips, takes part in resonating the laryngeal sounds to generate the voiced sounds with some bands of the formant frequencies, such as vowels. This is called *articulation*. The distribution pattern of the formants makes us discriminate the kind of vowel being produced. It is determined by the resonance property of the SVT, and the property in turn is dictated by the volumetric topology of the tract, which can be estimated from a function of the sequential cross-sectional area along the tract (Fant 1960; Stevens 1998; Titze 1994). The human SVT is composed of the almost equally long oral and pharyngeal cavities, and the epiglottis is separated from the velum, which produces a long oropharyngeal region facing the dorsal surface of the tongue, rostral to the laryngopharyngeal part that faces the epiglottis (Fig. 2; Crelin 1987; Laitman and Reidenberg 1993; Lieberman 1984). The tongue is globular to fit this configuration, and the internal musculature of the tongue makes the surface highly mobile (Takemoto 2001). The activities of the jaw and hyoid which provide a base for the tongue musculature have a major role in tongue movements (Hiiemae and Palmer 2003; Williams 1995; Zemlin 1988). Thus, the movements of the tongue, jaw, hyoid and lips are coordinated voluntarily to modify the SVT topology sequentially and rapidly; such a sophisticated regulation enables us to produce a series of distinct speech sounds even in a short single exhalation.

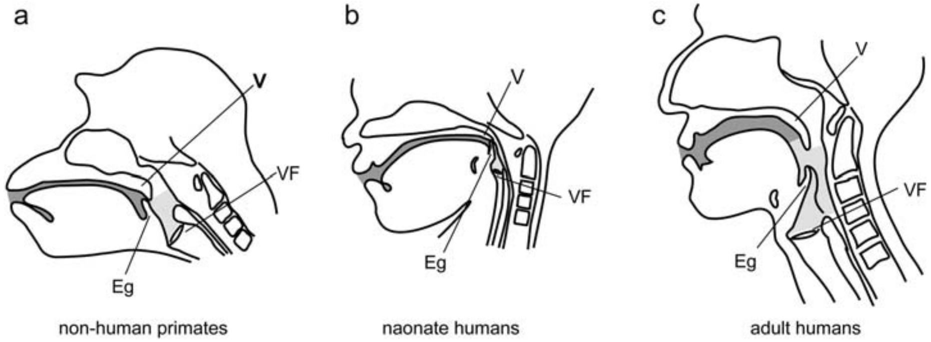


Fig. 2. Mid–sagittal diagrams of the head and neck in non-human primates, human neonates and adults. **a**, non-human primates. The pharyngeal cavity (colored in light grey) is much shorter and their epiglottis (*Eg*) is locked to the velum (*V*). The oral cavity (colored in dark grey) is much longer than the pharyngeal cavity. **b**, neonate humans. The pharyngeal configuration is similar to that in non-human primates with an intranarial position of the epiglottis. **c**, adult humans. The pharyngeal cavity is lengthened through the major descent of the larynx, and the supralaryngeal vocal tract is composed of the almost equally long oral and pharyngeal cavities. The epiglottis descntd from and lost the contact to the velum, to secure the long oropharyngeal region. *VF*, vocal fold (from Nishimura (2006) with permission).

3 Fossil Records

Paleoanthropologists have made great efforts to unveil the evolution of the physical foundations for human speech, based on the understanding of speech physiology. Efforts have mainly used two approaches: first, focusing on the issue of *voluntariness* in regulating the movements and activities of the vocal apparatuses; second, on the anatomical constraints for the sophisticated movements of the apparatuses. Although this discipline has examined various anatomical features that conceivably underlie these properties, different studies have claimed quite different ages for the origin of human speech: these estimates range over almost the whole history of the hominin lineage (Fig. 3). The following section surveys such efforts and arguments along with the various introduced assumptions and associated issues.

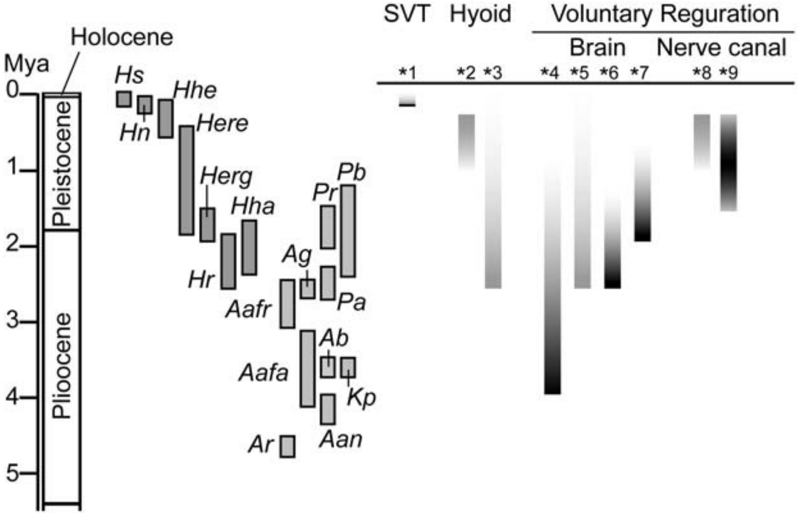


Fig. 3. Summary of the suggested ages of the origin of human speech and language. The estimates range over almost the whole history of the hominin lineage. abbreviations for hominins: *Aafa*, *Australopithecus afarensis*; *Aafr*, *A. africanus*; *Aan*, *A. anamensis*; *Ab*, *A. bahrelghazali*; *Ag*, *A. garhi*; *Ar*, *Ardipithecus ramidus*; *Here*, *Homo erectus*; *Herg*, *H. ergaster*; *Hha*, *H. habilis*; *Hhe*, *H. heidelbergensis*; *Hn*; *H. neanderthalensis*; *Hr*, *H. rudolfensis*; *Hs*, *H. sapiens*; *Kp*, *Kenyanthropus platyops*; *Pa*, *Paranthropus aethiopicus*; *Pb*, *P. boisei*; *Pr*, *P. robustus*.*[1] reconstruction of Neanderthal supralaryngeal vocal tract (SVT: Lieberman and Crelin 1971; Lieberman et al. 1972; cranial base: Laitman and Heimbuch 1982; Laitman et al. 1979).*[2] Neanderthal hyoid (Arensburg et al. 1989; 1990).*[3] australopithecine hyoid (based on Alemseged et al. 2006).*[4] Broca's area in the endocranial casts (Holloway 1976; 1983).*[5] endocranial feature in australopithecines (Falk 1980).*[6] endocranial feature in *H. habilis* (Tobias 1987).*[7] Brain volume (Walker 1993; Aiello and Dunbar 1993).*[8] size of the hypoglossal canal (Kay et al. 1998).*[9] size of the thoracic vertebral foramen (MacLarnon 1993; MacLarnon and Hewitt 1999).

3.1 Reconstructions of the Pharyngeal Anatomy

The SVT has a critical role in modulating the laryngeal sounds to produce distinct speech sounds; thus, fluent and rapid modifications of its topology are indispensable to speech production. Although this sophisticated manipulation is partly attributed to some neurological modifications underlying voluntary regulation of the movements of the peripheral anatomies, it is also believed to depend on modifications to the anatomies *per se* (Laitman and Reidenberg 1993; Lieberman 1984; Lieberman et al. 1969). Although most mammals, including humans, share the same basic anatomy of the SVT, nonhuman mammals show much shorter or smaller pharyngeal cavities wherein the epiglottis is locked to the velum and prevents the cavity from facing the movable tongue (Fig. 2: Fitch 2000a; Laitman and Reidenberg 1993; Negus 1949). Such a position of the epiglottis relative to the velum is called the *intranarial position*. Thus, in contrast to humans, they share an oral cavity that is long relative to the pharyngeal cavity, and the tongue is long in the horizontal direction (Fig. 2).

Lieberman et al. (1969) examined the potential to modify tongue posture and SVT topology in macaques and suggested that these primates probably produce a more restricted range of vowels than humans do. They argued that this deficit is probably attributable to their relatively short pharyngeal cavity and the intranarial position of the epiglottis. This argument followed many studies on reconstructions of the pharyngeal anatomy in fossil humans that attempted to deduce the likely position of the larynx and epiglottis relative to the palate. Lieberman and his colleagues examined the bony features of the cranial base, mandible and cervical vertebrae in the reconstructed skull of the Neanderthal specimen from La-Chapelle-aux-Saint to reconstruct its laryngeal position (Lieberman and Crelin 1971; Lieberman et al. 1972). They concluded that Neanderthals were unable to produce the full range of vowels produced by extant humans and argued that the full faculty for speech arose following the emergence of modern humans (Fig. 3). This argument was supported by comparative studies on the shape of the cranial base in extant humans, fossil humans and extant nonhuman primates. Laitman and his colleagues found that, in contrast to nonhuman primates, in extant humans the cranial base develops to be dominantly flexed in infancy, suggesting that such a developmental flexion is associated with the major descent of the hyoid and larynx seen in that period (Laitman and Heimbuch 1984; Laitman et al. 1978). Their statistical studies showed a significant distinction in the shape of the cranial base in modern humans from others, including Neanderthals, supporting the hypothesis of the origin of speech in modern humans (Fig. 3: Laitman and Heimbuch 1982; Laitman et al. 1979). On the other hand, various studies have criticized these conclusions, concerned mainly with the improper reconstruction of the Neanderthal skull *per se*, the less objective reconstruction of its laryngeal position, and underestimations in the potential for SVT manipulation (Boë et al. 2002; Carlisle and Siegel 1974; Duchin 1990; Falk 1975; Frayer 1993; Houghton 1993; Lieberman and McCarthy 1999).

Thus, although there have been many arguments for and against the accuracy of reconstructions of the pharyngeal anatomy in Neanderthals, they are all concerned with the potentials for SVT manipulation in this group: in other words, whether the full faculty of speech was shared by Neanderthals or arose solely in modern humans. This controversy has in part affected arguments on the classification of the Neanderthals as a subspecies of *Homo sapiens* or as an independent species, *Homo neanderthalensis*. (e.g., Klein 1989; Tattersall 1998).

3.2 Fossil Hyoid Bones

The hyoid bone provides a basis for the tongue's musculature and suspends the laryngeal skeleton caudally through muscles, ligaments and membranes. It is not articulated directly with the rest of the skeleton and is therefore called a *floating bone* in the neck (Fig. 4). It is the only bony element in the pharyngeal region and therefore can be fossilized and found in association with other human remains. There are four examples of fossilized hominin hyoids known: a set from Neanderthals, two pieces of Middle Pleistocene *Homo* and a part from Australopithecine specimen.

The first record is composed of most elements of the hyoid apparatus from Middle Paleolithic Neanderthal fossils found in Kebara Cave, Israel (Arensburg et al. 1989; 1990). Most features of the hyoid are available, and the bone is quite similar to that of modern humans in shape and dimension. This may support the argument that Neanderthals shared the low position of the hyoid and larynx as seen in modern humans, suggesting an earlier origin of the faculty for speech in a common ancestor (Fig. 3). However, hyoid morphology *per se* probably bears little relation to its position in the neck (Lieberman et al. 1989). Although the hyoid supports the coordinated cyclic activities of the tongue, there is a distinc-

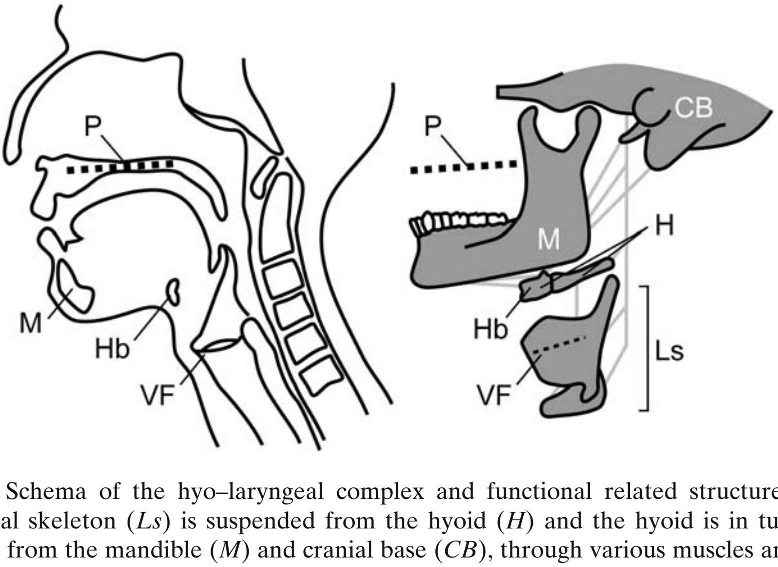


Fig. 4. Schema of the hyo–laryngeal complex and functional related structures. The laryngeal skeleton (*Ls*) is suspended from the hyoid (*H*) and the hyoid is in turn suspended from the mandible (*M*) and cranial base (*CB*), through various muscles and ligaments (grey lines). *Hb*, hyoid body; *P*, palate; *VF*, vocal folds (modified from Nishimura 2006 with permission).

tion in its activity between deglutition and speech (Hiiemae and Palmer 2003; Hiiemae et al. 2002). This suggests that any distinct feature of the Neanderthal hyoid may relate to such manipulations of the tongue's body in speech, rather than to its position.

The second record is composed of a well-preserved fossilized hyoid body from a three-year-old infant *Australopithecus afarensis*, from a Pliocene (3.3 Mya) deposit in the Dikika area of Ethiopia (Alemseged et al. 2006). This was bulla shaped, in contrast to the bar-like hyoid body in extant humans, suggesting that Australopithecines had a laryngeal air sac (Alemseged et al. 2006). While extant humans and gibbons share no true laryngeal sac, all great apes and siamangs share one that extends ventrally from bilateral laryngeal ventricles to form a centrally fused frontal sac (Hayama 1970; Hewitt et al. 2002; Negus 1949; Starck and Schneider 1960). In chimpanzees and gorillas, the fused sac expands superiorly to form an unpaired recess at the dorsal aspect of the hyoid body, and the bulla-shaped hyoid body receives this recess (Avril 1963; Raven 1950). In addition, this recess is formed at the latest at four months of age in chimpanzees (Nishimura et al. 2007). These facts strongly indicate that the Dikika baby had a laryngeal air sac, although its size is unknown. It should be noted that the Kebara evidence indicates that no laryngeal sac was present in Neanderthals, as is the case of extant humans. Two pieces of hyoids from Middle Pleistocene Homo hyderbergensis, Simade los Huesos in the Sierra de Atapuerca, Spain are also humanlike in both shape and size (Martínez et al. 2008). The functions of this sac remain matters of debate. Suggested functions include a storage of expired air to increase oxygen uptake (Negus 1949); reduction of the hyperventilation caused by a long sequence of repetitive loud calls (Hewitt et al. 2002); a generator of another sound source in the laryngeal ventricles (Brandes 1932; Fitch and Hauser 2003; Huber 1931; Kelemen 1948); a resonator of the laryngeal voice source to help produce loud and long calls (Fitch and Hauser 2003; Gautier 1971; Marler and Tenaza 1977; Napier and Napier 1985; Schön 1971; Schön Ybarra 1995), or as a buffer against the pressure induced by intensive expiratory airflow following air trapping during three-dimensional arboreal locomotion (Hayama 1970; 1996). These functions—excluding the first and last—are relevant to vocalization as often seen in the great apes, such as pant hoots in chimpanzees, suggesting that Australopithecines may have principally performed such vocalization but not speech-like utterances of a series of distinct sounds in a single exhalation (Fig. 3).

Thus, while hyoid morphology has played little contribution to the evaluation of the possible faculty of speech among Neanderthals and *Homo hyderbergensis* the hyoid of the Australopithecine baby might indicate chimpanzee-like vocalization rather than human speech. Clearly, hyoid morphology, without relation to the SVT reconstruction, needs to be examined further in living primates including humans to help reconstruct the likely tongue movements in fossil hominins.

3.3 Evaluation of the Voluntary Regulation of Vocalization

To produce speech sounds, humans show voluntary regulation of activities that are usually involuntary. Thus, the regulation of pulmonary activities is usually involuntary to allow safe tidal respiration. The hyolaryngeal apparatus and pharynx are also coordinated involuntarily during deglutition to propel boluses of food and liquid from the oropharynx to the esophagus, preventing them from incidental entrance to the trachea. The tongue moves less intentionally in feeding, although its activities vary slightly according to the physical nature of the foods and liquids being swallowed. However, humans are able to coordinate the activities of these apparatuses appropriately for speech production in a highly sophisticated manner. This faculty has probably followed evolutionary modifications to the neurological foundations for motor regulation.

Broca's area in the neocortex is known to contain the motor speech area that regulates the vocal apparatus to produce speech, and is distinct in extant humans. Brain tissue *per se* is not available for fossils, but many neurocranial fossils provide endocranial casts that permit rough evaluations of the brain surface structure in fossil hominins (Holloway 1974; 1983). Such studies suggest a region corresponding to the Broca's area in extant humans that have began to evolve in the South African *Australopithecus africanus* (Holloway 1976; 1983) and a distinct area in *Homo habilis* (Tobias 1987) , suggesting that these forms may have shared the neurological requirements for speech or language (Fig. 3). However, their reconstructions of the brain structure and their evaluations are still under disputation, and the other suggest a later or recent origin in hominin lineage (Fig. 3: Falk 1980). In addition, if a distinct evolutionary enlargement in brain volume *per se* is important for language, this may have arisen in the emergency of *Homo* excluding *Homo habilis* (Fig. 3: Aiello and Dunbar 1993; Walker 1993). However, it remains unclear that the modifications involve any functional modifications associated with speech. Moreover, it has yet to be determined which morphological features of endocranial casts reflect the true surface structure of the brain. Computed tomography enables us to nondestructively visualize and reconstruct endocranial features in fossil crania (e.g., Conroy et al. 1998), and such approaches may provide valuable information for future studies on the evolution of speech.

Aside from Broca's area, recent studies have concentrated on the relative size of the nerve canals through which motor fibers innervate the peripheral vocal apparatus. Such studies are based on the assumption that sophisticated regulation of the peripheral anatomy is associated with increased numbers of motor nerve fibers innervating it. Kay and his colleagues examined the hypoglossal canal in the cranial base, through which the hypoglossal nerves innervate the tongue musculature. They showed that early modern *Homo sapiens*, Neanderthals and middle Pleistocene *Homo* had canals falling in size within the range of modern humans, although Australopithecines and *Homo habilis* showed the canal falling within the range of the African apes (Kay et al. 1998). This supports arguments that human-like sophisticated activities of the tongue go back to the middle Pleistocene (>0.3 Mya: Fig. 3). However, this argument has been refuted by studies showing that the absolute and relative sizes of this canal overlap in extant hominoids, including humans (DeGusta et al. 1999; Jungers et al. 2003). In addition, this canal is traversed by other tissues, including veins of the hypoglossal plexus (Jungers et al. 2003; Williams 1995). This reduces any reliability in asserting that the size of this canal reflects that of the hypoglossal nerve (DeGusta et al. 1999; Jungers et al. 2003). Another line of study is concerned with the relative size of the thoracic vertebral foramen, from which the motor nerves innervate the pulmonary musculature regulating respiratory activities (MacLarnon 1993; MacLarnon and Hewitt 1999). These studies showed distinct enlargements of this canal in Neanderthals, falling within the range of extant humans, while the canal in the early Pleistocene *Homo ergaster* (1.6 Mya: Fig. 3) falls within the range of nonhuman hominoids. Thus, this feature appears to be associated with sophisticated regulation of respiratory activities, supporting arguments for an early origin of speech in the *Homo* lineage.

Thus, these approaches have all suggested an earlier origin of the faculty of speech among *Homo* including Neanderthals. However, this conclusion is based on the assumption that modifications to any neuroanatomical feature, including enlargements of the Broca's area or motor nerve canals, accompanied improvements in the physical activities underlying speech production. This assumption clearly needs further rigorous testing.

4 Dynamic Processes in Nonhuman Mammalian Vocalizations

Paleoanthropologists have searched for unique morphological features underlying human speech, including the peripheral anatomical features associated with sophisticated movements or the voluntary regulation of the vocal apparatus. However, the supposed uniqueness of extant humans has been challenged by increased knowledge of the relevant anatomy and bioacoustics in living nonhuman mammals. Such recent studies will provide information for our better understanding of the common ground of human speech and nonhuman vocalizations, and on the unique nature of the former. This section surveys these new disciplines, suggesting that integration of the results promises to advance our understanding of the evolution of human speech and language.

4.1 Descent of the Larynx: Is It Unique to Humans?

The hyoid and larynx are positioned at a low level relative to the palate in adult humans, and this feature had been believed to form the unique pharyngeal and SVT anatomy underlying speech production (Crelin 1987; Laitman and Reidenberg 1993; Lieberman 1984). In human neonates, these structures are positioned close to the palate, and the epiglottis is in the intranarial position (Fig. 2: Crelin 1987; Negus 1949; Zemlin 1988). They descend along the neck in the infant and early juvenile periods (Lieberman et al. 2001; Roche and Barkla 1965; Vorperian et al. 1999; 2005; Westhorpe 1987). This descent pulls the tongue down to the pharynx and makes the epiglottis descend relative to the velum, thereby lengthening the oropharyngeal region (Fitch and Giedd 1999; Sasaki et al. 1977; Vorperian et al. 1999; 2005). Nonhuman primates also show the descent of the hyoid and larynx relative to the palate, which lengthens the pharyngeal cavity during growth (e.g., macaque monkeys; Flügel and Rohen 1991). However, principally caused by technical limitations to access the region of interests, the major descent of the larynx as seen in humans had been believed to be lacking in nonhuman primates (Fig. 2: Laitman and Reidenberg 1993; Lieberman 1984; Negus 1949) and the term "descent" of the larynx or hyoid often specifically implies the proportional changes of the SVT toward the human configuration.

Recent studies using MRI have shown that chimpanzees also show a major descent of the larynx with the descent of the epiglottis from the velum, as in humans, despite their SVT configuration in the adult (Nishimura et al. 2003; 2006). On the other hand, adult faces grow to be prognathic to allow for a great elongation in the oral cavity and tongue in chimpanzees (Nishimura 2005; Nishimura et al. 2006). Although the developmental descent of the larynx and hyoid have been shown by a CT study in macaque monkeys (Flügel and Rohen 1991), the epiglottis maintains contact with the velum to restrict the pharyngeal cavity from facing the movable tongue even in adults (Flügel and Rohen 1991; Geist 1965; Laitman et al. 1977). Another anatomical study showed a difference in the spatial relationship of the laryngeal skeleton and hyoid in adults between hominoids and the other anthropoids (Nishimura 2003). Although uncertainty remains regarding the evolutionary path involving the major descent of the larynx before the divergence of the human and chimpanzee lineages, these findings suggest that the hominoid primates share the full descent process seen in humans and chimpanzees.

Even if the major descent of the larynx constitutes the "ontogeny" of the morphological foundations of speech in humans (Lieberman 1984), studies in nonhuman primates do not support the "evolutionary" hypothesis that this descent arose uniquely in the human lineage with a selective advantage for speech production. The developmental process is likely to have arisen before the diversification of the hominin lineage, probably in response to selection pressures unrelated to speech, and were secondarily advantageous for the development of human speech. The larynx and pharynx play important roles for the coordination of deglutition and breathing, suggesting that an advantage in such physiology may partly account for the initial selective advantages of the developmental processes (Nishimura 2003; 2006).

Thus, these achievements strongly suggest that, rather than using one unique feature as evidence for the evolution of speech and language, multidisciplinary and comparative approaches should contribute greatly to our understanding of the evolution of the biological foundations for speech and language, against phylogenetic and functional backgrounds.

4.2 Dynamic Manipulation of the Vocal Apparatus in Nonhuman Mammals

In nonhuman primates, the anatomy of the vocal apparatus had been believed to prevent the dynamic actions of SVT as seen in human speech and to restrict them to stereotyped and monotonous vocalizations. This argument principally depends on the knowledge on a static anatomy, namely the SVT configuration with long oral and short pharyngeal components in nonhuman primates, but not on the observations on the SVT activities in vocalizations. The vocal apparatus shows a unique anatomy, in that its frame is articulated with no other part of the bony skeleton (Williams 1995; Zemlin 1988)). Thus, anatomically the actions of the vocal apparatus are less constricted than the other components of the skeleton and cranium, and its potentials have been reevaluated in nonhuman mammals including primates.

Both cineradiography and X-ray television have been used for analyses of the action of the SVT in humans (e.g., Fant 1960), and recent improvements in digital imaging techniques have contributed to the study of human speech physiology (Story 2005; Story et al. 2001; Takemoto et al. 2006). The former approaches contributed to our understanding that nonhuman mammals share the physical ability to produce the SVT configuration that is achieved in human speech. Nonhuman mammals, including dogs, goats, pigs and cottontop tamarins, show a temporary lowering of the larynx to separate the epiglottis from the velum during loud calls (Fitch 2000b). Moreover, the position of the larynx in the males of some species of deer is similar to that in adult humans; they also lower further the larynx to produce resonance at lower frequencies during the roaring of the rut season (Fitch and Reby 2001). Such a potential is supported by anatomical studies on the hyolaryngeal region and pharynx in some carnivores, suggesting a dynamic lowering of the larynx in aspects of vocalization (Weissengruber et al. 2002). In fact, Diana monkeys produce a wide range of formant patterns and formant transitions, which had been believed to be unique to human speech (Riede and Zuberbühler 2003). This suggests that such vocalizations are probably produced by sequential and fluent modifications of the configuration of the SVT, including the temporal lowering of the larynx (Riede et al. 2005). Such a non-uniform shaping of the SVT may be found in the varied body size of domestic dogs (Riede and Fitch 1999).These studies suggest that the nonhuman mammals share an SVT that is more flexible than believed previously.

These recent advances also shed light on the arguments on the evolutionary advantage of the descended larynx. Reby et al. (2005) argued that lower frequencies in the roaring of red deer stags, which are produced by the temporary lowering of the larynx (Fitch and Reby 2001), possibly contribute to exaggeration of their perceived body size, to increase the repulsive effects against other males during the rutting season. Diana monkeys use their sophisticated vocal signals for predator-specific alarms (Riede and Zuberbühler 2003). These findings suggest that the selective advantage or advantages of vocalization, independent of speech and language, could account for the static low position of the larynx and physical ability of the dynamic actions of the SVT as seen in humans.

Thus, such dynamic capabilities of the SVT in nonhuman mammals strongly indicate that critical evidence for the unique nature in peripheral activities underlying speech production will be provided by the evaluation of dynamic activities to perform physiological functions, not just by evaluation of the static anatomy of the SVT.

4.3 Kinematic Approaches to the Study of Primate and Human Vocalizations

Nonhuman mammals, including primates, can show temporary lowering of the larynx and probably shape a non-uniform of SVT, to permit a wide range of acoustic formants and some degree of formant transitions. Nevertheless, they do show a distinction in the static anatomy of the SVT compared with humans. Morphologists and paleoanthropologists might have overemphasized this distinction and hence underestimated the ability of nonhuman primates and fossil humans to manipulate the SVT. However, static anatomical constrictions have limited influences on temporary changes in the posture of the vocal apparatus and SVT, but do constrain their entire activities. In fact, although nonhuman mammals show temporary lowering of the larynx, they show little in the way of further sophisticated activities in such a position (Fitch 2000b). By contrast, in human speech, the hyoid is shifted rostral to the region used in feeding to widen the pharyngeal cavity; it moves steadily and rapidly to produce a fluent sequence of speech sounds, although it moves in a region that is more constrained than during deglutition (Hiiemae and Palmer 2003; Hiiemae et al. 2002). Although this distinction may in part depend on neurological factors to regulate such movements, the permanent low position of the larynx and hyoid could provide the basis for less effortful, sophisticated and rapid rearrangement of the laryngeal position and tongue gesture (Fitch 2000b).

To evaluate this issue, kinematic or kinetic evaluations are expected to shed light on the anatomical restrictions to the dynamic activities of the SVT in primates, including humans. Although nonhuman primates share a potential for acoustic performance similar to human speech, such a performance has yet to be fully explained in physiological terms (Lieberman et al. 2006; Riede et al. 2006). For example, as for habitual bipedal walking, kinematic and kinetic studies have provided valuable information on the anatomical constrictions of this physical performance, with detailed examinations on the locomotion in living primates or simulations calculating the kinetic requisites (e.g., Kondo 1985; Ogihara and Yamazaki 2006; Tuttle 1981). These studies strongly suggest that the evolution of habitual bipedal walking was inevitably accompanied by anatomical modifications to the skeleton, and the knowledge has contributed to evaluations the locomotion forms and evolution in hominin lineage (e.g., Bramble and Lieberman 2004; Crompton et al. 1998; Senut et al. 2001; Wolpoff 1983). In fact, macaques trained for habitual bipedal walking over eleven years show compensatory modifications to the postcranial skeleton, including the vertebrae, within genetic limits (Hayama et al. 1992; Nakatsukasa et al. 1995). Such approaches have yet to be made in studies on speech and vocalizations, but must require many improvements in the information underlying the approaches, such as more detailed examinations of the anatomy of vocal apparatus and empirical studies on activities vocal apparatuses along with acoustics in nonhuman primates. Thus, the methodological and theoretical advances in bioacoustic studies have been increasing our knowledge on both topics as surveyed above. The integration of the two disciplines based on kinematic or kinetic analyses promise better understanding of the physiological uniqueness of human speech, which may have been overemphasized or underevaluated to date.

5 Summary

A brief explanation of speech physiology was offered here and it underlies any arguments about the distinctions between speech in humans and vocalization patterns in nonhuman primates. Human speech shows highly sophisticated activities of the SVT configuration through voluntary regulation of the vocal apparatuses, which are usually regulated involuntarily in other mammals. Comparative evaluations of these features in human speech and non-human vocalizations are a necessary part of any study on the evolution of language.

In the second section, the morphological and paleoanthropological efforts were surveyed. They have reconstructed the evolution of human speech, using reconstructions of the SVT and evaluations of the hyoid bone and nerve canals in fossil hominins. Although these studies used distinct morphological features, which are presumed to underlie the faculty of speech, the age estimated by the various studies covers almost the whole historical range of the human lineage.

In the third section, recent advances were reviewed in studies on the development of the anatomy of the SVT and on its dynamic manipulation, especially temporary lowering of the larynx, in nonhuman mammals. Although these investigations have contributed to our better understanding of these issues, the information needs to be integrated in kinematic or kinetic terms. Such integration promises to pave the way to reevaluate the unique nature of the physiology of human speech in the light of vocalizations in nonhuman mammals. It could contribute greatly to exploring the origin and evolution of the faculty of speech in paleoanthropological terms.

This paper has set out recent empirical studies on the anatomy and the dynamic activities of the vocal apparatus in nonhuman primates. I trust that it will stimulate further studies integrating both disciplines, to further our understanding of the evolution of human speech and language in biological terms.

Acknowledgements. This work was supported by a Research Fellowship of JSPS for Young Scientists (grant 16000326) and by the Global Center of Excellence Program "Formation of a Strategic Base for Biodiversity and Evolutionary Research: from Genome to Ecosystem" of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan (grant A06, to Kyoto University). I gratefully thank for the comments of the editor and my colleagues on the earlier version of the manuscript.

References

Aiello LC, Dunbar RIM (1993) Neocortex size, group size, and the evolution of language. Current Anthropology 34:184–193

Alemseged Z, Spoor F, Kimbel WH, Bobe R, Geraads D, Reed D, Wynn JG (2006) A juvenile early hominin skeleton from Dikika, Ethiopia. Nature 443:296–301

Arensburg B, Tiller AM, Vandermeersch B, Duday H, Schepartz LA, Rak Y (1989) A middle Palaeolithic human hyoid bone. Nature 338:758–760

Arensburg B, Schepartz LA, Tillier AM, Vandermeersch B, Rak Y (1990) A reappraisal of the anatomical basis for speech in Middle Palaeolithic hominids. American Journal of Physical Anthropology 83:137–146

Avril C (1963) Kehlkopf und kehlsack des Schimpansen, *Pan troglodytes* (Blumenbach 1799). Gegenbaurs morphologisches Jahrbuch 105:74–129

Boë L-J, Heim J-L, Honda K, Maeda S (2002) The potential Neandertal vowel space was as large as that of modern humans. Journal of Phonology 30:465–484

Bramble DM, Lieberman DE (2004) Endurance running and the evolution of *Homo*. Nature 432:345–352

Brandes R (1932) Über den kehlkopf des Orang-Utan in verschiedenen altersstadien mit besonderer berücksichtigung der kehlsackfrage. Gegenbaurs morphologisches Jahrbuch 69:1–61

Carlisle RC, Siegel MI (1974) Some problems in the interpretation of Neandelthal speech capabilities: A reply to Lieberman. American Anthropologist 76:9–325

Chiba T, Kajiyama M (1942) The vowel: Its nature and structure. Tokyo Kaiseikan Publishing, Tokyo

Christiansen MH, Kirby S (eds) (2003) Language evolution. Oxford University Press, New York

Conroy GC, Weber GW, Seidler H, Tobias PV, Kane A, Brunsden B (1998) Endocranial capacity in an early hominid cranium from Sterkfontein, South Africa. Science 280:1730–1731

Crelin ES (1987) The human vocal tract. Vantage Press, New York

Crompton RH, Li Y, Wang WJ, Gunther M, Savage R (1998) The mechanical effectiveness of erect and "bent-hip, bent-knee" bipedal walking in *Australopithecus afarensis*. Journal of Human Evolution 35:55–74

DeGusta D, Gilbert WH, Turner SP (1999) Hypoglossal canal size and hominid speech. Proceedings of the National Academy of Sciences of the United States of America 96:1800–1804

Duchin LE (1990) The evolution of articulate speech: Comparative anatomy of the oral cavity in *Pan* and *Homo*. Journal of Human Evolution 19:687–397

Falk D (1975) Comparative anatomy of the larynx in man and the chimpanzee: Implications for language in Neanderthal. American Journal of Physical Anthropology 43:123–132

Falk D (1980) A reanalysis of the South African australopithecine natural endocasts. American Journal of Physical Anthropology 53:525–539

Fant G (1960) Acoustic theory of speech production. Mouton, The Hague

Fitch WT (1997) Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. Journal of the Acoustical Society of America 102:1213–1222

Fitch WT (2000a) The evolution of speech: A comparative review. Trends in Cognitive Sciences 4:258–267

Fitch WT (2000b) The phonetic potential of nonhuman vocal tracts: Comparative cineradiographic observations of vocalizing animals. Phonetica 57:205–218

Fitch WT, Giedd J (1999) Morphology and development of the human vocal tract: A study using magnetic resonance imaging. Journal of the Acoustical Society of America 106:1511–1522

Fitch WT, Hauser MD (2003) Unpacking "honesty": Vertebrate vocal production and the evolution of acoustic signals. In: Simmons AM, Popper AN, Fay RR (eds) Acoustic communication. Springer–Verlag, New York, pp 65–137

Fitch WT, Reby D (2001) The descended larynx is not uniquely human. Proceedings of the Royal Society of London, Series B, Biological Sciences 268:1669–1675

Flügel C, Rohen JW (1991) The craniofacial proportions and laryngeal position in monkeys and man of different ages (A morphometric study based on CT-scans and radiographs). Mechanisms of Ageing and Development 61:65–83

Frayer DW (1993) On Neanderthal crania and speech: Response to Lieberman. Current Anthropology 34:721

Gautier JP (1971) Etude morphologique et fonctionnelle des annexes extra-laryngées des Cercopithecinae; liaison avec les cris d'espacement. Biologia Gabonica 7:229–267

Geist FD (1965) Nasal cavity, larynx, mouth, and pharynx. In: Hartman CG, Straus WL Jr (eds) The anatomy of the rhesus monkey. Hafner Publishing, New York, pp 189–209

Hayama S (1970) The *Saccus laryngis* in primates. Journal of the Anthropological Society of Nippon 78:274–298 (in Japanese with English abstract)

Hayama S (1996) Why does not the monkey fall from a tree?: The functional origin of the human glottis. Primate Research 12:179–206 (in Japanese with English summary)

Hayama S, Nakatsukasa M, Kunimatsu Y (1992) Monkey performance: The development of bipedalism in trained Japanese monkeys. Acta Anatomica Nipponica 67:169–185

Hewitt G, MacLarnon A, Jones KE (2002) The functions of laryngeal air sac in primates: A new hypothesis. Folia Primatologica 73:70–94

Hiiemae KM, Palmer JB (2003) Tongue movements in feeding and speech. Critical Reviews in Oral Biology and Medicine 14:413–429

Hiiemae KM, Palmer JB, Medicis SW, Hegener J, Jackson BS, Lieberman DE (2002) Hyoid and tongue surface movements in specking and eating. Archives of Oral Biology 47:11–27

Holloway RL (1974) The casts of fossil hominoid brains. Scientific American 231:106–115

Holloway RL (1976) Paleoneurological evidence for language origins. Annals of the New York Academy of Sciences 280:330–348

Holloway RL (1983) Human paleontological evidence relevant to language behavior. Human Neurobiology 2:105–114

Houghton P (1993) Neandertal supralaryngeal vocal tract. American Journal of Physical Anthropology 90:139–146

Huber E (1931) Evolution of facial musculature and expression. The Johns Hopkins Press, Baltimore, MD

Jungers WL, Pokempner AA, Kay RF, Cartmill M (2003) Hypoglossal canal size in living hominoids and the evolution of human speech. Human Biology 75:473–484

Kay RF, Cartmill M, Michelle B (1998) The hypoglossal canal and the origin of human vocal behavior. Proceedings of the National Academy of Sciences of the United States of America 95:5417–5419

Kelemen G (1948) The anatomical basis of phonation in the chimpanzee. Journal of Morphology 82:229–256

Klein RG (1989) The human career: Human biology and cultural origins. The University of Chicago Press, Chicago, IL

Kondo S (ed) (1985) Primate morphophysiology, locomotor analyses and human bipedalism. The University of Tokyo Press, Tokyo

Laitman JT, Heimbuch RC (1982) The basicranium of plio-pleistocene hominids as an indicator of their upper respiratory systems. American Journal of Physical Anthropology 59:323–343

Laitman JT, Heimbuch RC (1984). A measure of basicranial flexion in Pan paniscus, the pygmy chimpanzees. In: Susman RL (ed) The pygmy chimpanzee. Plenum, New York, pp 49–63

Laitman JT, Reidenberg JS (1993) Specialization of the human upper respiratory and upper digestive system as seen through comparative and developmental anatomy. Dysphasia 8:318–325

Laitman JT, Crelin ES, Conlogue GJ (1977) The function of the epiglottis in monkey and man. Yale Journal of Biological Medicine 50:43–48

Laitman JT, Heimbuch RC, Crelin ES (1978) Developmental change in a basicranial line and its relationship to the upper respiratory system in living primates. American Journal of Anatomy 152:467–482

Laitman JT, Heimbuch RC, Crelin ES (1979) The basicranium of fossil hominids as an indicator of their upper respiratory system. American Journal of Physical Anthropology 51:15–34

Lieberman DE, McCarthy RC (1999) The ontogeny of cranial base angulation in humans and chimpanzees and its implications for reconstructing pharyngeal dimensions. Journal of Human Evolution 36:487–517

Lieberman DE, McCarthy RC, Hiiemae KM, Palmer JB (2001) Ontogeny of postnatal hyoid and larynx descent in humans. Achieves of Oral Biology 46:117–128

Lieberman P (1984) The biology and evolution of language. Harvard University Press, Cambridge, MA

Lieberman P (2006) Limits on tongue deformation—Diana monkey formants and the impossible vocal tract shapes proposed by Riede et al. (2005). Journal of Human Evolution 50:219–221

Lieberman P, Blumstein SE (1988) Speech physiology, speech perception, and acoustic phonetics. Harvard University Press, Cambridge, MA

Lieberman P, Crelin ES (1971) On the speech of Neanderthal man. Linguistic Inquiry 2:203–222

Lieberman P, Crelin ES, Klatt DH (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. American Anthropologist 74:287–307

Lieberman P, Laitman JT, Reidenberg JS, Landahl K, Gannon PJ (1989) Folk psychology and talking hyoids. Nature 342:486–487

Lieberman PH, Klatt DH, Wilson WH (1969) Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. Science 164:1185–1187

MacLarnon AM (1993) The vertebral canal. In: Walker A, Leakey R (eds) The Nariokotome *Homo erectus* skeleton. Harvard University Press, Cambridge, MA, pp 359–390

MacLarnon AM, Hewitt GP (1999) The evolution of human speech: The role of enhanced breathing control. American Journal of Physical Anthropology 109:341–363

MacLarnon AM, Hewitt GP (2004) Increased breathing control: Another factor in the evolution of human language. Evolutionary Anthropology 13:181–197

Marler P, Tenaza R (1977) Signaling behavior of apes with special reference to vocalization. In: Sebeok TA (ed) How animals communicate. Indiana University Press, Bloomington, IN, pp 965–1033

Martínez I, Arsuaga JL, Quam R, Carretero JM, Gracia A, Rodrigues L (2008) Human hyoid bones from the middle Pleistocene site of the Sima de los Huesos (Sierra de Atapuerca, Spain). Journal of Human Evolution 54:118–124

Nakatsukasa M, Hayama S, Preuschoft H (1995) Postcranial skeleton of a macaque trained for bipedal standing and walking and implications for functional adaptation. Folia Primatologica 64:1–29

Napier JR, Napier PH (1985) The natural history of the primates. MIT Press, Cambridge, MA

Negus V (1949) The comparative anatomy and physiology of the larynx. William Heinemann Medical Books, London

Nishimura T (2003) Comparative morphology of the hyo-laryngeal complex in anthropoids: Two steps in the evolution of the descent of the larynx. Primates 44:41–49

Nishimura T (2005) Developmental changes in the shape of the supralaryngeal vocal tract in chimpanzees. American Journal of Physical Anthropology 126:193–204

Nishimura T (2006) Descent of the larynx in chimpanzees: Mosaic and multiple-step evolution of the foundations for human speech. In: Matsuzawa T, Tomonaga M, Tanaka M (eds) Cognitive development in chimpanzees. Springer–Verlag, Tokyo, pp 75–95

Nishimura T, Mikami A, Suzuki J, Matsuzawa T (2003) Descent of the larynx in chimpanzee infants. Proceedings of National Academy of Sciences of the United States of America 100:6930–6933

Nishimura T, Mikami A, Suzuki J, Matsuzawa T (2006) Descent of the hyoid in chimpanzees: Evolution of facial flattening and speech. Journal of Human Evolution 51:244–254

Nishimura T, Mikami A, Suzuki J, Matsuzawa T (2007) Development of the laryngeal air sac in chimpanzees. International Journal of Primatology 28:483–492

Ogihara N, Yamazaki N (2006) Computer simulation of bipedal locomotion: Toward elucidating correlations among musculoskeletal morphology, energetics, and the origin of bipedalism. In: Ishida H, Tuttle R, Pickford M, Ogihara N, Nakatsukasa M (eds) Human origins and environmental backgrounds. Springer, New York, pp 167–174

Raphael LJ, Borden GJ, Harris KS (2007) Speech science primer: Physiology, acoustics, and perception of speech, 5th edn. Lippincott Williams & Wilkins, Baltimore, MD

Raven HC (1950) The anatomy of the gorilla. Columbia University Press, New York

Reby D, McComb K, Cargnelutti B, Darwin C, Fitch WT, Clutton–Brock T (2005) Red deer stags use formants as assessment cues during intrasexual agonistic interactions. Proceedings of the Royal Society of London, Series B, Biological Sciences 272:941–947

Riede T, Fitch T (1999) Vocal tract length and acoustics of vocalization in the domestic dog (*Canis familiaris*). Journal of Experimental Biology 202:1859–1867

Riede T, Zuberbühler K (2003) The relationship between the acoustic structure and semantic relationship in Diana monkey alarm vocalization. Journal of the Acoustical Society of America 114:1132–1142

Riede T, Bronson E, Hatzikirou H, Zuberbühler K (2005) Vocal production mechanisms in a non-human primate: Morphological data and a model. Journal of Human Evolution 48:85–96

Riede T, Bronson E, Hatzikirou H, Zuberbühler K (2006) Multiple discountinuities in nonhuman vocal tracts—A response to Lieberman (2006). Journal of Human Evolution 50:222–225

Roche AF, Barkla DH (1965) The level of the larynx during childhood. Annals of Otology, Rhinology and Laryngology 74:645–654

Sasaki CT, Levine PA, Laitman JT, Crelin ES (1977) Postnatal descent of the epiglottis in man: A preliminary report. Archives of Otolaryngology 103:169–171

Schön MA (1971) The anatomy of the resonating mechanism in howling monkeys. Folia Primatologica 15:117–132

Schön Ybarra MA (1995) A comparative approach to the non-human primate vocal tract: Implications for sound production. In: Zimmerman E, Newman JD, Jürgens U (eds) Current topics in primate vocal communication. Plenum Press, New York, pp 185–198

Senut B, Pickford M, Gommery D, Mein P, Cheboi K, Coppens Y (2001) First hominid from the Miocene (Lukeino Formation, Kenya). Comptes Rendus de l'Académie des Sciences. Series IIA, Earth and Planetary Sciences 332:137–144

Starck D, Schneider R (1960) Respirationsorgane: A. Larynx. In: Hofer H, Schultz AH, Starck D (eds) Primatologia, vol 3-2. Karger, Basel, pp 423–587

Stevens KN (1998) Acoustic phonetics. MIT Press, Cambridge, MA

Story BH (2005) A parametric model of the vocal tract area function for vowel and consonant simulation. Journal of the Acoustical Society of America 117:3231–3254

Story BH, Titze IR, Hoffman EA (2001) The relationship of vocal tract shape to three voice qualities. Journal of the Acoustical Society of America 109:1651–1667

Takemoto H (2001) Morphological analyses of the human tongue musculature for three-dimensional modeling. Journal of Speech, Language, and Hearing Research 44:95–107

Takemoto H, Honda K, Masaki S, Shimada Y, Fujimoto I (2006) Measurement of the temporal changes in vocal tract area function from 3D cine-MRI data. Journal of the Acoustical Society of America 119:1037–1049

Tattersall I (1998) Becoming human: Evolution and human uniqueness. Harcourt Brace, New York

Tuttle RH (1981) Evolution of hominid bipedalism and prehensile capabilities. Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences 292:89–94

Titze IR (1994) Principles of voice production. Prentice Hall, Englewood Cliffs, NJ

Tobias PV (1987) The brain of Homo habilis: A new level of organization in cerebral evolution. Journal of Human Evolution 16:741–761

Vorperian HK, Kent RD, Gentry LR, Yandell BS (1999) Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: Preliminary results. International Journal of Pediatric Otorhinolaryngology 49:197–206

Vorperian HK, Kent RD, Lindstrom MJ, Kalina CM, Gentry LR, Yandell BS (2005) Development of vocal tract length during early childhood: A magnetic resonance imaging study. Journal of the Acoustical Society of America 117:338–350

Walker A (1993) Perspectives on the Nariokotome discovery. In: Walker A, Leakey R (eds) The Nariokotome *Homo erectus* skeleton. Harvard University Press, Cambridge, MA, pp 411–430

Weissengruber GE, Forstenpointner G, Peters G, Kubber-Heiss A, Fitch WT (2002) Hyoid apparatus and pharynx in the lion (*Panthera leo*), jaguar (*Panthera onca*), tiger (*Panthera tigris*), cheetah (*Acinonyxjubatus*) and domestic cat (*Felis silvestris f. catus*). Journal of Anatomy 201:195–209

Westhorpe RN (1987) The position of the larynx in children and its relationship to ease of intubation. Anaesthesia and Intensive Care 15:384–388

Williams PL (ed) (1995) Gray's anatomy, 38th edn. Churchill Livingstone, New York

Wolpoff MH (1983) Lucy lower-limbs—Long enough for lucy to be fully bipedal. Nature 304:59–61

Wray A (ed) (2002) The transaction to language. Oxford University Press, New York

Zemlin WR (1988) Speech and hearing science: Anatomy and physiology, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ

9
Implication of the Human Musical Faculty for Evolution of Language

Nobuo Masataka

1 Introduction

Although 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 system arose by means of natural selection, presumably because more accurate communication helped early humans survive and reproduce. However, with respect to music, even the evolutionary significance remains open to question. Pinker (1997) stated that music itself played no adaptive role in human evolution, suggesting 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 and Kirby 2003; Wallin et al. 2000), arguing that music clearly had an evolutionary role, and pointing to music's universality. Particularly from a developmental perspective, findings concerning the ability of very young infants to respond strongly to music should be noted because it could serve as evidence that music is hardwired into human brains. Thus, if infants' musical ability is the result of Darwinian natural selection, in what way did it make humans more fit?

Concerning the evolution of language as a system, numerous findings have recently been reported in the area of cognitive science, particularly developmental cognitive science, providing material for presentation of a concrete scenario (for a review see Masataka 2003; 2007). These findings indicate that children learn the subcomponents of this system one after another during acquisition, 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. Nonetheless, linguists tend to take evolutionary onset as an abrupt phenomenon (e.g., Pinker 1994). On the other hand, the view of developmental cognitive scientists is that

Primate Research Institute, Kyoto University, Inuyama, Aichi 484-8506, Japan

the attainment of each subcomponent is intertwined in humans. In line with this, questions arise as to how each subcomponent evolved into its present form and what the previous forms were as well as when and how several subcomponents came to be related with one another. To answer each of these questions would help reconstruct language evolution, and accordingly, allow elucidation of the presence of the puzzling cognitive musical ability of infants. The present review is a preliminary attempt to provide a conceptual perspective on the above questions in an aim to help unravel the implications of the human musical faculty, particularly with reference to the evolution of language.

2 "Coo"s of Human Infants and of Japanese Macaque Adults

Of various early language developmental milestones, the earliest is perceptual competence, which is related to speech sounds in typically-developing infants exposed to spoken language (Werker and Voutoumanos 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 speechlike 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 vowellike monosyllabic sounds ("coo"s) at 6–8 weeks of age. The learning process that preverbal infants must undertaken 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 so-called "coo" calls to maintain vocal contact. When my colleague and I (Sugiura and Masataka 1995) observed vocal exchanges of the vocalizations in free-ranging populations, the temporal patterns of occurrence of intercall intervals between two consecutive coos during vocal interaction are similar to those obtained in human mother-infant dyad; when a monkey utters a coo spontaneously, the monkey remains silent for a short interval, and when no response occurs, the monkey is likely to give further coos addressing 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 these macaque vocalizations undergo similar modification as a function of social context. This evidence is summarized into 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 et al. 2006). When the frequency 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 underlying modification of a fixed template, which is similar to but more subtle than what has been reported in 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 receipt of distant calls.

As for 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 this study, similar phenomena have been reported in New World monkeys and great apes (Masataka 2003). In both these studies, flexible variability occurred acoustically in the vocalizations with respect to the pattern of temporal organization of 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, not in interactions involving juveniles or infants. Although longitudinal data is still required, some kind of social experience seems to be necessary for acquisition of this ability.

3 Prosodic Communication in Human Infants

In humans, on the other hand, similar abilities are already observed in preverbal infants before they are able to produce well-articulated sounds. Halliday (1975) made the first systematic attempt to examine the significance of within-individual variability of 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 to 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 and McLeod 1989). The primary role is referred to as the "attention-getting property". Even newborns exhibit a strong tendency to direct their attention 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 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 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 become more co-ordinated, 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.

4 Evolution of "Singing" Behavior

The onset of the second stage in the early process of spoken language production approximates to the age of 8 months when speechlike vocalizations in infancy culminates in the sense that babbling typically emerges. 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 adultlike spectral and temporal properties. One month later, it comes to be articulated with a rapid formant transition duration in a relatively short syllable. Results of acoustical various analyses with the vocalizations demonstrated that there is a significant continuity between the sound system of babbling and early speech, and that units present in babbling are utilized later in natural spoken languages (for a review see Masataka 2003). Accordingly, approximately one year after birth, first words are observed.

In nonhuman primate, no vocalizations are as well-articulated as babbling of 9-month-old human infants. Any form of multisyllabic utterance does not 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 multisyllabic sounds human infants as young as 8 months of age produce in a poorly-articulated manner. In particular, they have been investigated in gibbons most intensively. Darwin (1871) 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 analysis 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 male 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 and Seibt 1982).

Furthermore, male solo as well as duet song bouts were found to occur in mated pairs of one species (Geissmann 2002). The majority of these solo songs are heard approximately 2 hours after dawn but approximately 2 hours earlier than duet songs. In another group, however, the first peak of singing activity occurs at or even before sunrise, and this time, the males produce solo songs from their sleeping tree. The second peak occurs one to several hours later, after the first feeding bout then the females 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 the pattern of duetting is the most elaborate form 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 was split into temporally segregated solo songs.

5 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 actually regardless of the culture in which they live (Boulez 1971). It is therefore possible that 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.

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 upon such common cognitive properties, human infants are enabled to acquire spoken languages, and the properties have been predisposed to the infants genetically during the human evolution. Consequently they exhibit a finer pattern of discrimination when music stimuli are presented (Masataka 2006; Trainor and Heinmiller 1998; Trainor et al. 2002). Although it has been argued that the early pattern of language discrimination and recognition reflects innate language-specific predisposition unique to humans, this assumption has recently been challenged (Fishman et al. 2001). While in humans the neural mechanism for language processing is believed to be located in Wernicke's area, there is evidence for the fact that the fixation of a gene (*FOXP2*) expressing in Broca'a area occurred during the last 200,000 years of human history (Enard et al. 2002). Deficits of the gene is associated with the occurrence of motor speech disorders (Watkins et al. 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. On the other hand, 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 five to six 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, 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.

6 Experimental Evidence for the Volitional Control of Vocal Production in Gibbons

According to the hypothesis described here, 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 other parts, the result will no longer be singing. Moreover, if this executed under voluntary motor control and the influence of auditory feedback, the result could almost be construed as speech-like behavior. This possibility has been examined experimentally by us more recently (Koda et al. in submission). In the study, we attempted operant conditioning of the vocalizations of an immature female white-handed gibbon (*Hylobates lar*) housed at Fukuchiyama City Zoo, Kyoto Prefecture, Japan, which is approximately 120 km northwest of Kyoto. The animal, born from a mated pair that was also kept at the zoo, was 2 years and 6 months old when the study began.

At the age of 1 year and 2 months, the subject was separated from her parents. Since that time, she has been housed individually in an indoor cage located in a different building from that of her parents. The indoor cage was connected to an open enclosure area with a radius of approximately 3 m, surrounded by a fence. The subject usually spent the daytime in this open area. Two or three 0.7-m-high decks were separately located in the open area, where we conducted all the experiments.

The entire experiment consisted of six stages. The first three stages were the preliminary and preparation periods for the subsequent conditioning attempt. The latter three stages were more directly involved in the learning process of the training. All interactions between the subject and experimenter during the course of the training were recorded by a video camera with a microphone, which was located approximately 2 m away.

6.1 Stage I. Familiarization Stage

The purpose of this first stage was to familiarize the subject with the experimenter. For approximately 1 week, an experimenter frequently played with the subject as the animal sat on a deck in the outdoor enclosure. Here, "playing" refers to embracing, dodging, pulling, and tickling the animal. While playing with the subject, the experimenter was always seated in the same position and always attempted to keep the subject seated on a deck. At the end of this period, the subject learned to remain seated in front of the experimenter.

6.2 Stage II. Response-Soliciting Stage

For the following 2-week period, the experimenter repeated simple training in which she held a small piece of apple or banana in the right hand, closed the hand tightly, and extended the right arm toward the subject. Although the subject obviously attempted to take the food from the experimenter's hand, the experimenter never opened her hand unless the subject vocalized. After any type of vocalization was produced, the subject was immediately allowed access to the food. Unless the subject vocalized, the trial continued until the subject lost interest in taking the food reward or became fussy and moved away from the deck. Such trials of simple training were conducted opportunistically whenever the subject appeared cooperative; approximately 300 trials were intensively conducted during this period.

6.3 Stage III. Habituating Stage with the V-sign

The V-sign was simultaneously shown with the right hand after stage II, while a piece of apple or banana held in the left hand was shown to the subject, to let the subject associate the V-sign cue and her vocalization behavior. Otherwise, the protocol of this stage was exactly the same as that of the previous stage. Trials were conducted opportunistically and were intensively continued for 1 week.

6.4 Stage IV. Nonselective Conditioning Stage

The general protocol for the training in this stage was the same as that for stage III, and we conducted an experimentally planned test session to check the proficiency of the training. As the training proceeded, we performed a 10-trial test session as follows. A session began with a several-minute play period. A trial began with a face-to-face posturing of the subject seated on the deck, as in the previous stages. When the experimenter confirmed that the subject was calm and the experimenter had adequately attracted the subject's attention, the V-sign was displayed with the right hand held in front of the chest (Fig. 1). In a trial, the subject was required to produce any type of vocalization to obtain a piece of apple or banana from the experimenter. The subject could obtain the food reward immediately if she successfully vocalized. The V-sign was presented to the subject until the subject vocalized. If the subject did not produce any type of vocalization within 60 s, the experimenter did not provide a reward, and the next trial proceeded. Vocalizations were recorded if the subject vocalized within 60 s after the onset of the V-signing. When the response was recorded, its latency to

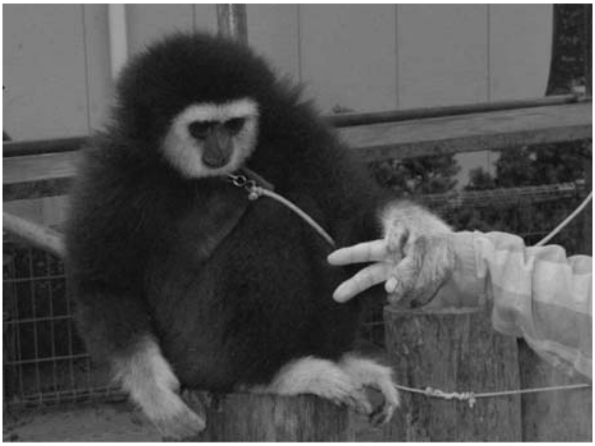


Fig. 1. Photograph of conditioning. The experimenter shows a V-sign cue shaped by the right hand in front of the subject. In the photo, the gibbon successfully vocalized in response to the V-sign and opened her mouth.

the onset of the signing was measured, with an accuracy of 0.1 s, by viewing the video recording. The intertrial interval was >10 s, and a single session was usually finished within 5 minutes. A maximum of two sessions were conducted in a single day. In total, 14 ten-trial sessions were conducted over 13 days.

6.5 Stage V. Selective Conditioning Stage

Contrary to the previous stage, this stage aimed to condition well-phonated calls. The training protocol of the previous stage elicited any type of vocalization, resulting in poorly phonated calls such as squeaks, screeches, and whines rather than well-phonated sounds designated as basic notes of the gibbon song (Raemaekers et al. 1984). However, as the main purpose of this study was to examine the behavioral foundations establishing complex songs, we examined whether well-phonated calls could be exclusively conditioned. In this stage, the protocol was essentially the same as that in the preceding stage except that the subject was rewarded only if she produced a well-phonated call, as judged by the experimenter. These calls were articulated so well that the experimenter could easily distinguish by ear between well-phonated and poorly phonated calls. In this stage, we conducted a 10-trial test session during the training to assess the proficiency of the training. This stage, with a total of 17 ten-trial test sessions, was conducted over 13 days. A maximum of two sessions were conducted in a single day.

6.6 Stage VI. Reconditioning Stage

After the completion of the conditioning of well-phonated vocalizations, 18 ten-trial test sessions of the same training protocol as for stage V were conducted after a 30-day interval to examine the effect of the interval on the performance of the subject. No training procedures involving vocal conditioning were conducted during these 30 days. All sessions were carried out for 3 days, and the maximum number of sessions conducted in a single day was limited to three.

To examine the successive course of the vocal conditioning, we focused on the performance of 10-trial test sessions during the latter three stages, i.e., stages IV, V, and VI. For the analysis, we measured the latency between the onset time of V-sign cueing and the subject's vocalizations. Next, we treated the trial where the subject vocalized within 10 s after the onset of the V-signing as the successful trials. We measured the performance of successful trials in the test sessions. To assess the performance of a particular session, we scored successful and failed trials as binomial data (success = 1, failure = 0). Moreover, we counted the number of incontingent vocalizations emitted during a 10-trial session. All measurements were confirmed by viewing the video recoding.

For the statistical analysis, we used a generalized linear model (GLM) with session as an explanatory continuous variable, with an appropriate error and link function. For binomial performance data, we used the binomial family and logit link function; for latency data, we used the Gaussian family and log link function; and for vocalization count data, we used the quasi-Poisson family and log link function. To examine the effect of session, we compared the predicted model to the null model, eliminating the session term from the predicted model by analysis of deviance using chi-square values for binomial and Gaussian families or F-values for the quasi-Poisson family (Crawley 2002). The GLM procedure was performed for each stage, using R software (R.2.1.0, R Development Core Team, http://www.R-project.org).

7 Successful Vocal Operant Conditioning in an Immature Gibbon

Figures 2–4 represent the percentages of successful trials, mean latency, and cumulative number of vocalizations in a session for stages IV, V, and VI, respectively. We generally found an increase in the percent success and decreases in the latency and cumulative number of vocalizations for each stage. The percentage of successful trials in the first half of the test sessions ranged from 40 to 100%, whereas the scores in the latter half were consistently and significantly higher at >80%. Although longer latencies were found in the early sessions, the latency became shorter at the end of the sessions. The decrease in the latency was significant and corresponded to the increase in training performance. The number of vocalizations emitted in a session was not stable; the numbers in the final three sessions were 5, 1, and 0. However, we found no significant decrease in the number of vocalizations as the sessions proceeded. The percentages of successful trials in the first half of the test sessions (20 to 90%) were lower than those in the latter half (consistently >70%), except in session 12 (Fig. 3). The increase in the percentage of successful trials was significant. Except in session 12, the laten-

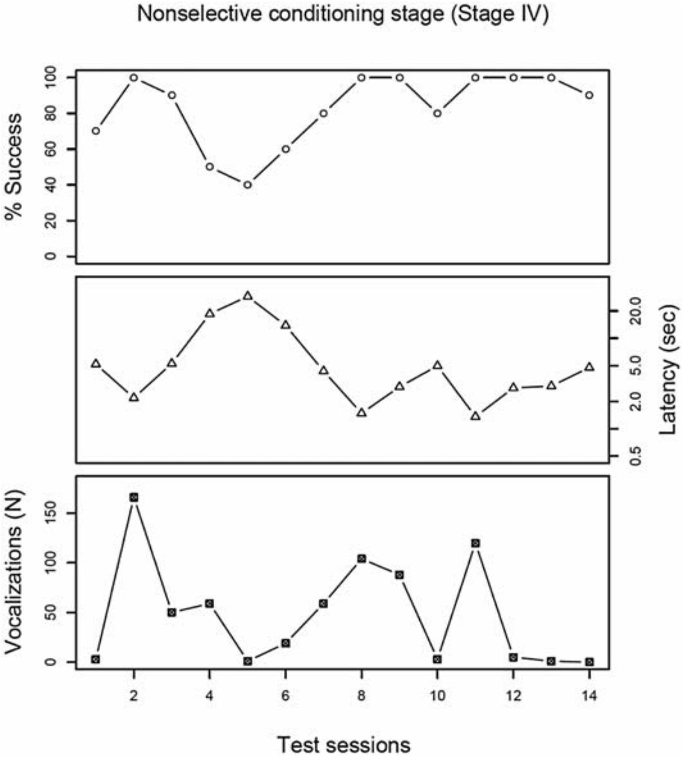


Fig. 2. Conditioning process in the nonselective conditioning stage (stage IV). (**Top**) Percentage of successful trials in 14 ten-trial sessions. (**Center**) Mean latencies of 14 ten-trial sessions. (**Bottom**) Cumulative number of vocalizations emitted per session in 14 ten-trial sessions (modified from Koda et al. 2007, with permission).

cies gradually yet significantly decreased as the test sessions proceeded, corresponding to the increase in training performance. There was no significant decrease in the number of vocalizations as the sessions proceeded, but the numbers of vocalizations were extremely low in all test sessions (Fig. 3). The maximum number of vocalizations was 9, occurring in the seventh test session.

A typical learning process was observed in this stage. In the first three sessions, the percentages of successful trials were 70, 30, and 40%. After the fourth test session, the percentages were consistently higher, always >80% (Fig. 4). The change in performance was so drastic that a statistical procedure was unnecessary to show significance. As the sessions proceeded, the latencies gradually yet significantly decreased, corresponding to enhanced training performance, and the number of vocalizations significantly decreased.

Generally, lower performances were observed in the early sessions of each stage, and consistently higher or perfect performances for the training trials were

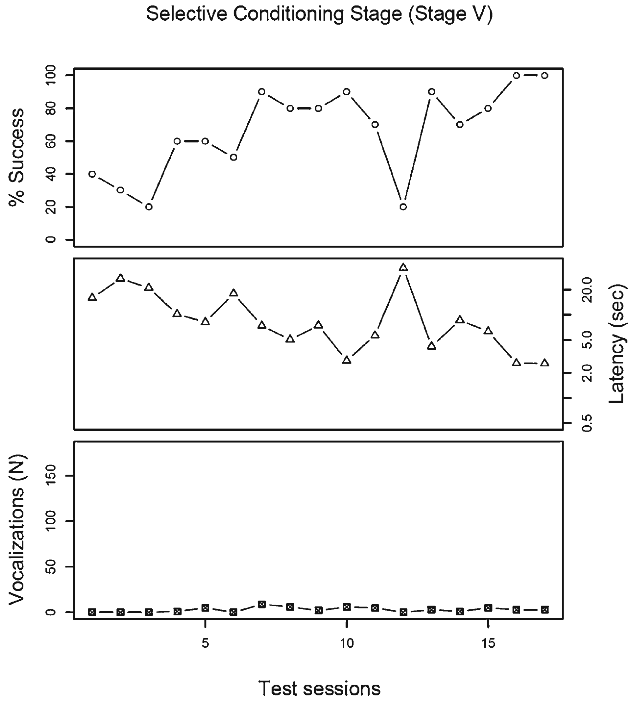


Fig. 3. Conditioning process in the selective conditioning stage (stage V). (**Top**) Percentage of successful trials in 17 ten-trial sessions. (**Center**) Mean latencies of 17 ten-trial sessions. (**Bottom**) Cumulative number of vocalizations emitted per session in 17 ten-trial sessions (modified from Koda et al. 2007, with permission).

confirmed in the later sessions of each stage. The latencies gradually decreased as sessions proceeded for each stage, corresponding to the improved performance results. The incontingent vocalizations were likely to be inhibited gradually as the subject learned to emit the appropriate vocalizations. The inhibition of incontingent vocalizations was supported by the extremely low frequency of vocalizations in stage V and an immediate decrease in the number of vocalizations in stage VI. The learning process in the reconditioning stage (stage VI) showed a typical pattern of conditioning. At the start of reconditioning, lower performance, longer latency, and greater frequency of vocalization were observed. These patterns were caused by the effect of the 30-day interval, which likely lead to an extension of the conditioning. However, the performance immediately recovered to that of the previous stage (V), at which the subject had been successfully conditioned to vocalize. These results suggest the successful conditioning of vocal production in this subject and provide direct evidence for the volitional control of vocal production.

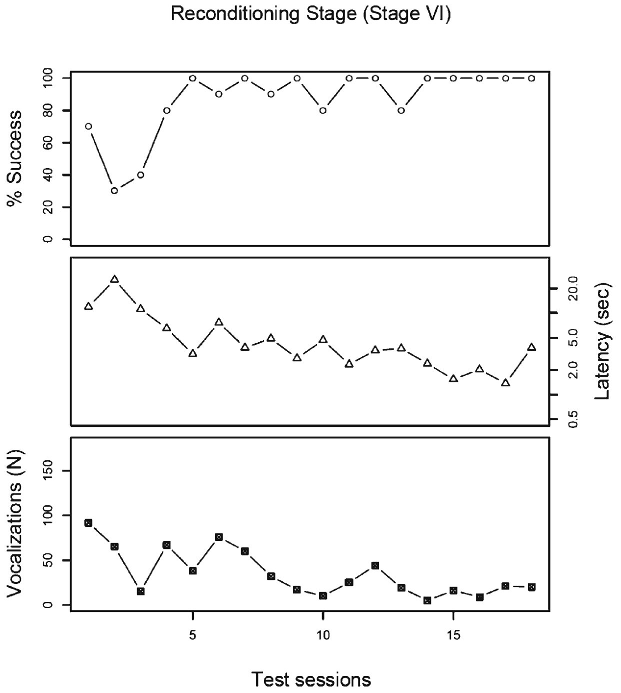


Fig. 4. Conditioning process in the reconditioning stage (stage VI). (**Top**) Percentage of successful trials in 18 ten-trial sessions. (**Center**) Mean latencies of 18 ten-trial sessions. (**Bottom**) Cumulative number of vocalizations emitted per session in 18 ten-trial sessions (modified from Koda et al. 2007, with permission).

8 Social Significance for Language-Like Vocal Behavior in Gibbons

Why has so elaborated vocal communication system evolved in gibbons? To answer the question would provide us with cues to solve the problem of language evolution.

The functions of songs and duets have been proposed as maintaining spatial organization among neighboring family groups and advertising territory to nonmate conspecifics to deter them from intruding upon an occupied territory (Haimoff 1984; Whitten 1982). These presumable functions of territorial advertisement and strengthening of pair bond in their duet songs were consistent with the songbird studies in the previous ethological literature (e.g., Farabaugh 1982). Taken together with variable aspects of specializations mentioned above, the traditional view of the social structure of gibbons had been proposed as a stable monogamous social structure, living in nuclear family groups (e.g., Leighton 1987).

However, some evidence against the traditional view of the rigid pattern of their monogamy has been reported. Palombit (1994) was the first to challenge the nuclear family model in hylobatids. He summarized a 6-year history of group changes in white-handed gibbons, and siamang, *Sympharangus syndactylus*, at Ketambe Research Station, Sumatra. Mate desertion and repairing were surprisingly frequent in the six study groups, and gave rise to groups that were not nuclear families. He concluded that "the complexities of social life in these animals extend beyond the narrow limits established by a rigid nuclear family concept". Subsequently Brockelman et al. (1998) firstly reported the demography data of the several wild white-handed gibbon groups in Khao Yai National Park, central Thailand based on the 18-years longitudinal observation. Their observations provided the surprising findings on reproduction, natal dispersal, pair formation, and group structure. They observed two cases in which the resident male in a group was replaced by a dispersing adult in the neighboring group. In one case, forcible displacement of the resident male resulted in a group which included a young juvenile presumable fathered by the previous male, two younger juveniles (probably brothers) from the new male's original group, and (later) offspring of the new pair. Nevertheless, social relations within this heterogeneous group appeared harmonious. Other works also suggested the possibility of the flexible social structure (Bartlett 2003; Fuentes 2000; Reichard 1995; Sommer and Reichard 2000). Recent results of DNA analysis on the faecal samples collected from wild gibbons in Khao Yai confirmed the flexible social structure supported by the wild observations, and showed evidence for the frequent occurrences of extra-pair copulation (Barelli et al. 2006). The rethinking of the "monogamous" social structure in gibbon species has been required (Sommer and Reichard 2000).

Recently we reported an observation on change of membership of a group of wild agile gibbons, *Hylobates agilis agilis*, living in west Sumatra (Koda et al. 2007). The wild agile gibbon in Sumatra Island has rarely been previously investigated. We observed possible replacement of the resident male by an adult male coming outside the home range on a group intensively investigated. Moreover, we examined the effect of the mate replacement on the territorial boundary changes, using the auditory census technique. This observation will contributes to the rethinking of the traditional view of the gibbon social system. In the study, we investigated a population of wild agile gibbons in a tropical rainforest of Andalas University in Padang (0°54′ S, 100°28′ E), west Sumatra, Indonesia, which consisted of a mixture of primary and secondary forest. Research took place from 12 September–31 November 2005 (first observational period) and 12 July–8 September 2006 (second observational period). Our previous investigation identified more than seven gibbon groups inhabited the study area, and their approximate home ranges were estimated by the direct observation (Oyakawa et al. 2007). The previous study reported marked individual acoustical differences in singing of gibbons under observation. In the present report, we focused on the B group, because we could comprehend all of the membership during the two periods in this group. The home range of the B group was located at the center of our research field.

During the two observations, we confirmed the changing of the membership of the B group. In the first observational period, we identified the five gibbons, which were one adult female (named as Gula), one adult male (Laki-laki), one subadult male (Malam), one subadult female (Bunga) and one infant male (Air). The Gula and Laki-laki were supposed to be pair mate gibbons of the B group at that time, because we usually heard and observed their organizing duets. On the other hand, we identified the three gibbons, which were Gula, Air and the new adult male (Galam) in the second observational period. Gula and Air were perfectly identified between the two periods, because this mother-infant pair possesses the significant physical character of the fur color. Their fur colors are bright buff, although the common is brown in the agile gibbon. Their identification must be successful. Whereas the fur colors of Laki-laki, Malam and Galam were bright brown, black and dark brown. Although we cannot deny the misidentification between Malam and Galam, we could no doubt confirm the replacement of the pair male between the two periods. Moreover, Galam is probably the different gibbon from Malam, because the body size is quite different. Malam was estimated as the subadult or adolescent male by his small body size, whereas Galam was already matured from his appearance of bigger body size. It was reasonable to treat as the different individuals because of the fur color and body size.

On the basis of the occurrence of the pair male replacement, we estimated the B group's home ranges of the two periods. In the first and second observational periods, the 29 and 49 singing location points were recorded, and the average numbers per a day were 1.16 and 2.04, respectively. In the first and second periods, the MCP estimations of the home range calculated 11.47 and 12.10 ha for the convex polygon dimension, respectively. The overlap dimension of the two convex polygons was 9.10 ha, and the percentage of the overlap area per the first period home range area was 79.3% (Fig. 5). The distance of the two gravity centers of 2005 and 2006 were calculated as 43.8 m.

9 Concluding Remarks

Our observation of the male replacement is consistent with several recent findings about gibbon social structure (Brokelman et al. 1998; Palombit 1994; Sommer and Reichard 2000), accumulating the evidences of the flexible and dynamic social structure of gibbons' "monogamous" systems. Moreover, we found the stability of the territorial boundary even when the mate replacement occurred in the subject group. Of particular interest should be the fact that no essential change of the territorial boundary occurred, and that a new mated pair took over the territory where a previous resident pair had defended. The effect of the

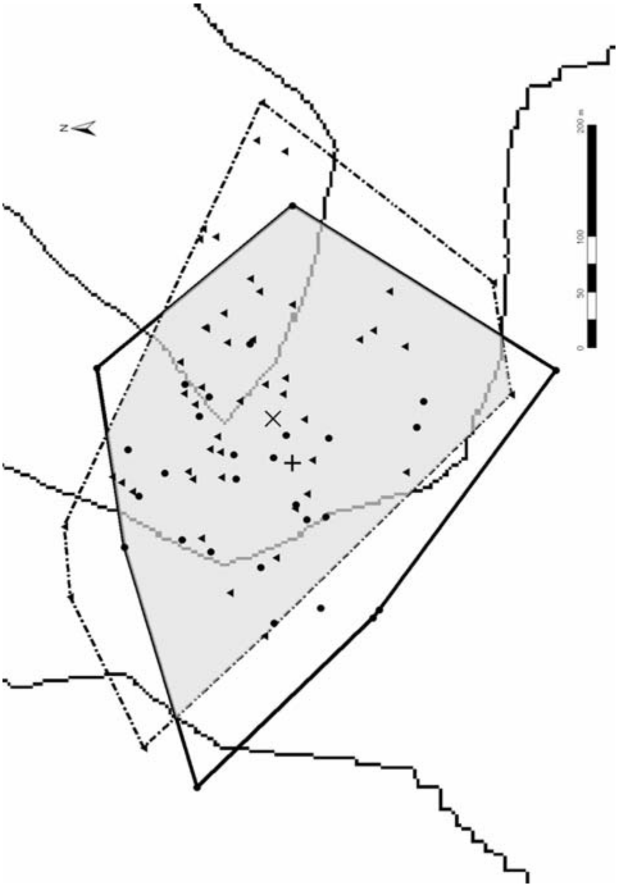


Fig. 5. Map of the study area. Closed circles, the singing locations of B group in 2005; Closed triangles, the singing locations of B group in 2006; Curved lines, 50-m contours (range: 250–350 m); black line polygon, minimum convex polygon (MCP) estimated as the home rage of B group in 2005; dotted line polygon, MCP estimated as the home rage of B group in 2006; gray area, overlapping area between the two home ranges of 2005 and 2006; *+ mark*, gravity center of MCP in 2005; *× mark*, gravity center of MCP in 2006.

membership change of gibbon groups on their territory has been rarely examined except for the longitudinal study of white-handed gibbon (Brokelman et al. 1998). They reported that the territory size increased or decreased, corresponding with the replacement of the pair mate males or numbers of the group member. During the replacement of the pair male, serious wounding was observed in an adult male of the mated pair of the neighboring group. The fact should be noticeable that he no longer duetted with his pair mate female regularly, had difficulty in keeping up with his group, and finally disappeared. Thereafter, three male offspring of the pair attempted to replace an adult male living in a group adjacent to their own group and succeed with it. One of the three male achieved the pair replacement as the forcible take-over of group, the other two males dispersed and organized the new groups in adjacent areas. Although the dynamic changes of the group organization were observed above, the territorial boundary was not virtually affected by such social alteration. Singing of resident gibbons is strongly associated with the stability of territories where they reside.

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. Recent studies with gibbon "singing" behavior indicate the possibility that our unique human faculty for music may be a distinct mental module, from which language has evolved. It should be noted, moreover, that there may be a common theme running through our music capacity and gibbon singing capacity. They are self-expressive. They cost time and energy, but neither of them has no clear survival benefits. They show strong individual differences, require experiences, and health. They play upon the perceptual and cognitive preferences of spectators. These are all the hallmarks of adaptations that have been shaped as courtship ornaments by Darwin's process of sexual selection through mate choice. Taken together, the above findings may serve as evidence for emergence of language by humans as evolution of cultural displays as production of music and production of gibbon singing do.

References

Barelli C, Heltermann M, Hodges KJ, Boesch C, Reichard U (2006) Female sexual activity in relation to reproductive endocrinology and mating systems in white-handed gibbons (*Hylobates lar*). Paper presented at the 21st Congress of the International Society of Primatology

Bartlett TQ (2003) Intragroup and intergroup social interactions in white-handed gibbons. International Journal of Primatology 24:239–259

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

Brockelman WY, Reichard U, Treesucon U, Raemaekers JJ (1998) Dispersal, pair formation and social structure in gibbons (*Hylobates lar*). Behavioral Ecology and Sociobiology 42:329–339

Christiansen MH, Kirby S (2003) Language evolution. Oxford University Press, Oxford

Crawley M (2002) Statistical computing: An introduction to data analysis using S-Plus. John Wiley & Sons, Chichester

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

Dunbar RIM (1993) Coevoluion of neocortex size, group size and language in humans. Behavioral and Brain Sciences 16:681–735

Enard W, Przeworski W, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, 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

Farabaugh E (1982) The ecological and social significance of duetting. In: Kroodsma D, Miller E, Ouellet H (eds) Acoustic communication in birds. Academic Press, New York, pp 85–124

Fishman YI, Volkov IO, Noh MD, Garell PC, Bakken H, Arezzo JC, Howard MA, Steinschneider M (2001) Consonance and dissonance of musical chords: Neural correlates in auditory cortex of monkeys and humans. Journal of Neurophysiology 86:2761–2788

Fuentes A (2000) Hylobatid communities: Changing views on pair bonding and social organization in hominoids. Yearbook of Physical Anthropology 43:33–60

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

Haimoff E (1984) Acoustic and organizational features of gibbon song. In: Preuschoft H, Chivers D, Brockelman W, Creel N (eds) The Lesser apes: Evolutionary and behavioural biology. Edinburgh University Press, Edinburgh, pp 333–353

Haimoff EH (1986) Convergence in the duetting of monogamous Old World primates. Journal of Human Evolution 15:51–59

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

Koda H, Oyakawa C, Kamilah SN, Hayakawa S, Chaniago RB, Suigiura H, Masataka N (in submission) Male replacement and stability of territorial boundary in a group of agile gibbons (*Hylobatis agilis agilis*) in West Sumatra, Indonesia

Koda H, Oyakawa C, Kato A, Masataka N (2007) Experimental evidence for the volitional control of vocal production in an immature gibbon. Behaviour 144:681–692

Leighton D (1987) Gibbons: Territoriality and monogamy. In: Smuts B, Cheney D, Seyfarth R, Wrangham R, Struhsaker T (eds) Primate societies. University of Chicago Press, Chicago, pp 134–187

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

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

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 (2007) Music, evolution and language. Developmental Science 10:35–39

Oyakawa C, Koda H, Sugiura H (2007) Acoustic features contributing to the individuality of wild agile gibbon (*Hylobates agilis agilis*) songs. American Journal of Primatology

Palombit R (1994) Dynamic pair bonds in hylobatids: Implications regarding monogamous social system. Behaviour 128:65–101

Pinker S (1994) The language instinct. Harper Collins, New York

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

Raemaekers JJ, Raemaekers PM (1984) Loud calls of the gibbon (*Hylobates lar*): repertoire, organization and context. Behaviour 91:146–189

Reichard U (1995) Extra-pair copulations in a monogamous gibbon (*Hylobates lar*). Ehology 100:99–112

Sommer V, Reichard U (2000) Rethinking monogamy: The gibbon case. In: Kappeler P (ed) Primate males: Causes and consequences of variation in group composition. Cambridge University Press, Cambridge, pp 254–298

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: Zimmermann E, Newman JD, Jurgens U (eds) Current topics in primatology. Plenum, London, pp 121–140

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 LJ, Heinmiller BM (1998) The development of evaluative responses to music: Infants prefer to listen to consonance over dissonance. Infant Behavior and Development 21:77–88

Trainor LJ, Tsang CD, Cheung VHW (2002) Preference for sensory consonance in 2- and 4-month-old infants. Music Perception 20:187–194

Wallin NL, Merker B, Brown S (2000) The origins of music. MIT Press, Cambridge, MA

Watkins KE, Dronkers NF, Vergha-Khadem F (2002) Behaviral analysis of an inherited speech and language disorder: Comparison with acquired aphasia. Brain 125:452–464

Werker JF, McLeod PJ (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 JF, Voutoumanos A (2000) Who's got rhythm? Science 288:280–281

Whitten A (1982) The ecology of singing in Kloss gibbons (*Hylobates klossii*) on Siberut Island, Indonesia. International Journal of Primatology 3:33–51

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

Subject Index