IV THEORIES OF MUSIC ORIGIN

16 The “Musilanguage” Model of Music Evolution

Steven Brown

Abstract

Analysis of the phrase structure and phonological properties of musical and linguistic utterances suggests that music and language evolved from a common ancestor, something I refer to as the “musilanguage” stage. In this view, the many structural features shared between music and language are the result of their emergence from a joint evolutionary precursor rather than from fortuitous parallelism or from one function begetting the other. Music and language are seen as reciprocal specializations of a dual-natured referential emotive communicative precursor, whereby music emphasizes sound as emotive meaning and language emphasizes sound as referential meaning. The musilanguage stage must have at least three properties for it to qualify as both a precursor and scaffold for the evolution of music and language: lexical tone, combinatorial phrase formation, and expressive phrasing mechanisms.

Beyond Music-Language Metaphors

Theories of music origin come in two basic varieties: structural models and functional models. Structural models look to the acoustic properties of music as outgrowths of homologous precursor functions, whereas functional models look to the adaptive roles of music as determinants of its structural design features. This chapter presents a structural model of music evolution. Functional models are presented elsewhere (Brown in press).

Before discussing music from an evolutionary perspective, it is important to note that two different modes of perceiving, producing, and responding to musical sound patterns exist, one involving emotive meaning and the other involving referential meaning. These I call, respectively, the acoustic and vehicle modes. The acoustic mode refers to the immediate, on-line, emotive aspect of sound perception and production. It deals with the emotive interpretation of musical sound patterns through two processes that I call “sound emotion” and “sentic modulation.” It is an inextricably acoustic mode of operation. The vehicle mode refers to the off-line, referential form of sound perception and production. It is a representational mode of music operation that results from the influence of human linguistic capacity on music cognition.[1] The vehicle mode includes the contexts of musical performance and contents of musical works, where both of these involve complex systems of cultural meaning (see footnote 2 for details).

This distinction between the acoustic and vehicle modes addresses an important issue in contemporary musicology: the conflict between absolutists, who view music as pure sound-emotion, and referentialists, who view it as pure sound-reference (discussed in Feld and Fox 1994). Seeing music in terms of the acoustic mode-vehicle mode duality permits reconciliation of the two viewpoints by suggesting that two different modes of perceiving, producing and responding to musical sound patterns exist, one involving emotive meaning and one referential meaning. These two modes act in parallel and are alternative interpretations of the same acoustic stimulus.

The very notion of a vehicle mode for music (or of referentialism) leads immediately to the question of the extent to which music functions like a language. Serious consideration of this question dates back at least to the eighteenth century if not earlier (Thomas 1995). No doubt the question hinges on the criteria by which one calls a given system a language, and this has led many thinkers to clarify notions of musical syntax and semantics (Bernstein 1976; Sloboda 1985; Clarke 1989; Aiello 1994; Swain 1995, 1996). The reciprocal question deals with the extent to which speech exploits musical properties for the purposes of linguistic communication in the form of speech melody and rhythm. But, whereas the metaphors go both ways, from language to music and back again, it is important to realize that these accounts are only ever seen as metaphors. Concepts such as musical language (Swain 1997) and speech melody are never taken beyond the domain of metaphor into the domain of mechanism. That is why, to me, this metaphor making misses the point that music and language have strong underlying biological similarities in addition to equally strong differences. Converging evidence from several lines of investigation reveals that the similarities between music and language are not just the stuff of metaphors but a reflection of something much deeper.

Given the extensive practice of metaphor making in linguistics and musicology, how can we best think about the similarities that exist between music and language? (I discuss only the acoustic route of language communication, and thus speech. A discussion of gesture, which is relevant to the evolution of both language and dance, will be presented at a future time.) Perhaps the best place to start is at the point of greatest distinction: grammar. The grammar metaphor is quite pervasive in musicology. The notion that musical phrase structures (can) have a hierarchical organization similar to that of linguistic sentences, an idea presented elegantly by Lerdahl and Jackendoff (1983), must be viewed as pure parallelism. In other words, the hierarchical organization of pitches and pulses in a Bach chorale is only loosely related to the hierarchical organization of words in a sentence exactly because the constituent elements, and thus the phrases themselves, are so completely different. However, to the extent that the generativity analogy works at all in music, it is only because of important underlying features (which Lerdahl and Jackendoff themselves make mention of in their closing pages) that provide a biological justification for this potential for hierarchical organization in music. What this means is that music and language must converge at some deep level to have hierarchical organization flower from two such different grammatical systems.

What is this point of convergence? The answer, briefly, is combinatorial syntax and intonational phrasing. First, in both language and music, the phrase is the basic unit of structure and function. It is what makes speaking and singing different from grunting and screaming. In both, a limited repertoire of discrete units is chosen out of an infinite number of possible acoustic elements, such that phrases are generated through combinatorial arrangements of these unitary elements. Thus, the use of discrete building blocks and the generation of higher-order structures through combinatorial rules is a major point of similarity between music and language. But it is not the whole story, as both make extensive use of expressive phrasing. Phrasing refers to modulation of the basic acoustic properties of combinatorially organized phrases for the purposes of conveying emphasis, emotional state, and emotive meaning. It can occur at two levels, local and global. Local modulation selectively affects individual elements of the phrase in the context of the whole phrase, whereas global modulation affects the whole phrase in a rather equivalent manner. From this standpoint, both speech phrases and musical phrases are melodorhythmic structures in which melody and rhythm are derived from three sources: acoustic properties of the fundamental units (pitch sets, intensity values and duration values in music; phonemes and phonological feet in speech); sequential arrangement of such units in a given phrase (combinatorial rules in both domains); and expressive phrasing mechanisms that modulate the basic acoustic properties of the phrase for expressive emphasis and intention (phrasing rules in both domains).

These properties of combinatorial syntax and intonational phrasing set the stage for the overall structural features of music and language. Perhaps the most important realization about their cognitive organization is that both systems function on two separate levels, and that these levels emerge out of the common set of principles described above (figure 16.1). One plane is the phonological level and the other is the meaning level. The first one is acoustic and is based on the principles of discreteness, combinatoriality, and phrasing. It is governed by a type of phonological syntax (see Marler, this volume) dealing with the selection and organization of sound units for the purposes of communication. The meaning level is where these acoustic elements are interpreted for higher-order signification in a context-dependent and cultural fashion. It is here that we see the greatest divergence between music and language, as the elements of the phonological level feed into very different systems of meaning. In language, phonological units are interpreted as lexical words, and are fed into a system of propositional syntax, which can be used to describe the properties of objects or express an ordered set of relationships between actors and those acted upon. It can express relationships about being, intention, causality, possession, relatedness, history, and so on. In music's acoustic mode, the units of the phonological level are interpreted as motivic, harmonic, rhythmic, and timbral elements, and are fed into a system of pitch-blending syntax that specifies a set of relationships between sound patterns and emotion. It deals with the issues of sound emotion, tension and relaxation, rhythmic pulse, and the like. Music's vehicle mode involves an interaction between these two syntax types, as described below.

Thus, both music and language consist of two related but dissociable tiers, each derived from a common set of principles dealing with phrases and phrasing. The end result of this analysis is the realization that phonological phrases and meaningful phrases are related but distinct entities. This fact is well known in linguistics, where the relationship between intonational phrases and syntactic phrases is at best probabilistic (Pierrehumbert 1991; Ladd 1996; Cruttenden 1997). It is no less true of music. However, the effect for language is much more striking from an evolutionary standpoint, as this liberation of language's meaning level from the acoustic modality (phonological level) allows language to develop into a system of amodal representation so important in theories of symbolic representation and off-line thinking (Bickerton 1995).

Five Possible Models

Space limitations prevent me from providing a general analysis of the phrase structure of music and language. My goal here will merely be to place this issue in an evolutionary perspective: How can we account for the similarities between music and language in evolutionary terms? Can we talk about mechanisms rather than metaphors? To this end, it will be important to distinguish two types of features that are shared between music and language: *shared ancestral* and *analogous* features, terms taken from the theory of cladistic classification in evolutionary biology. The first group have their roots in the common evolutionary origins of music and language. The second group arise due to the parallel but independent emergences of similar processes during the evolution of music and language. Aside from these shared ancestral and analogous features are the *distinct* features that are unique to either music or language.

To the extent that music and language share underlying phonological and syntactic properties, we can imagine five basic evolutionary possibilities by which this could have occurred (figure 16.2). First, these

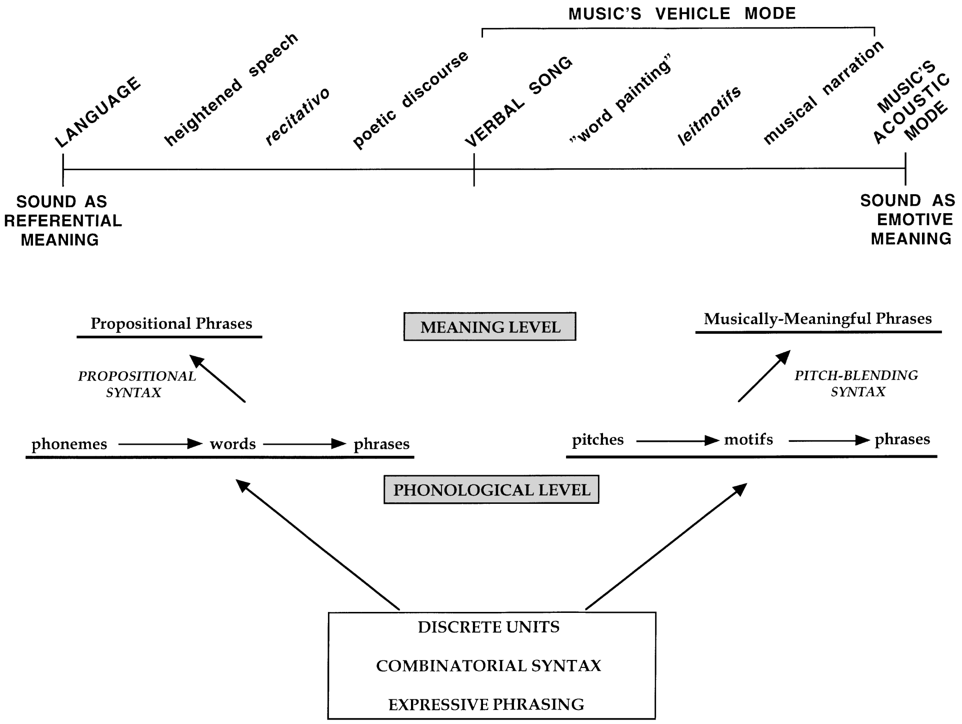

Figure 16.1
The two levels of functioning of music and language: phonological and meaning. Both levels are derived from the process of phrase formation involving discrete units, combinatorial syntax, and expressive phrasing. The phonological level is the acoustic level. It is governed by a type of phonological syntax in which discrete acoustic units (phonemes, pitches) are combined to form functional units (morpheme, motifs) that feed into the meaning level of each system. The meaning levels of language and music are governed by different types of syntax systems: propositional and blending, respectively. At their highest level of function, music and language differ more in emphasis than in kind, and this is represented by their placement at different ends of a spectrum. The poles of the spectrum represent the different interpretations of communicative sound patterns that each system exploits in creating meaningful formulas, where language emphasizes sound as referential meaning and music emphasizes sound as emotive meaning. A large number of functions occupy intermediate positions along this spectrum in that they incorporate both the referentiality of language and the sound-emotion function of music. Verbal song is the canonical intermediate function, which is why it occupies the central position. The functions of music's vehicle mode (see footnote 2 for details) lie toward the music side, whereas linguistic functions that incorporate sound-emotion or isometric rhythms lie toward the language side of the spectrum. ("Word painting" refers to the technique by which a composer creates an iconic relationship between music and words, such as the association of a descending melodic contour with the word "falling." This is use of music as symbolizer, as described in footnote 2).

Fig. 16.2

Figure 16.2
Five models for the evolution of the shared properties of music and language. In the parallelism model, language's evolution from a protolinguistic precursor and music's evolution from a protomusical precursor are thought to occur by completely independent processes. The binding model is quite similar except that it posits evolution of binding mechanisms that confer linguistic properties onto music and musical properties onto language (shown by the reciprocal horizontal arrows at the top of the figure). Neither of these two models invokes any notion of shared ancestral features. The next three models do. In the music outgrowth model, music is thought to evolve out of a linguistic precursor, whereas in the language outgrowth model language is thought to evolve out of a musical precursor. The musilanguage model is another outgrowth model in which shared properties of music and language are attributed to a common precursor, the musilanguage stage.

similarities could have come about completely fortuitously and arisen purely by parallel evolution. This parallelism model rejects any notion of shared ancestral features. Second, the similarities could have arisen from continuing interaction between discrete music and language modules, such that effective binding mechanisms evolved to confer musical properties onto language and linguistic properties onto music (binding model). Third, music could have evolved as an outgrowth of language (music outgrowth model). Fourth, language could have evolved as an outgrowth of music (language outgrowth model). Fifth, these similarities could have arisen due to the occurrence of an ancestral stage that was neither linguistic nor musical but that embodied the shared features of modern-day music and language, such that evolutionary divergence led to the formation of two distinct and specialized functions with retention of the shared features conferred onto them by the joint precursor (musilanguage model). Compared with the first two models, the last three invoke shared ancestral traits as being the basis for at least some similarities between music and language, but posit different evolutionary paths for their emergence.

I propose the musilanguage model for the origins of music and language. Why not adopt one of the other models? First, music and language have just too many important similarities for these to be chance occurrences alone. The parallelism model is the least parsimonious of the group evolutionarily. The binding model, which is implicitly the model of contemporary neurological studies (manifested by the credo "language: left hemisphere, music: right hemisphere"), rests on an overly dichotomous view of music and language, and is refuted by any type of neurological lesion that eliminates the musical properties of speech but spares those of music, or vice versa. Thus, studies showing that selective anesthesia of the right hemisphere of the brain disrupts the proper use of pitch during singing but leaves speech prosody intact (Borchgrevink 1991) indicate that binding models are too dichotomous. This is where outgrowth models present advantages. They assume that outgrowth of one function from the other permits not only the sharing of features due to common ancestry but redundant representation in the brain of similar functions by virtue of the divergence and differentiation events that led to outgrowth.

My reason for preferring the musilanguage model over either outgrowth model is that it greatly simplifies thinking about the origins of music and language. As it uses the common features of both as its starting point, the model avoids the endless semantic qualifications as to what constitutes an ancestral musical property versus what constitutes an ancestral linguistic property, exactly the kind of uncertainty that makes outgrowth models difficult to justify. The model forgoes this by saying that the common features of these two systems are neither musical nor linguistic but *musilinguistic*, and that these properties evolved first. In contrast, the distinct features of music and language, which are those that theorists can more or less agree upon, occurred evolutionarily later. They are specializations that evolved out of a common precursor and are thus (metaphorically) like the various digits that develop out of a common limb bud during ontogeny of the hand.

The model posits the existence of a musilanguage stage in the evolution of music and language (see figure 16.2). This stage must satisfy two important evolutionary criteria: first, it must provide for the common structural and expressive properties that are found in music and language (the shared ancestral features); and second, and quite important, it must provide an evolutionary scaffold on which music and language can evolve after a period a divergence and differentiation. In other words, the stage must be as much a precursor for the origins of language as it is for the origins of music, and should not have properties that are either too musical to permit evolution of language or too linguistic to permit evolution of music.

The Musilanguage Model

Much of what is described here was inspired by two basic ideas about music and language. The first one is the musilanguage idea, which contends that the two evolved as specializations from a common ancestral stage, such that their shared ancestral features evolved before their distinct, differentiated properties. The second idea is that despite the ultimate divergence between music and language during human evolution, these two functions differ more in emphasis than in kind, and are better represented as fitting along a spectrum instead of occupying two discrete, but partly overlapping, universes (see the top of figure 16.1). At one end of this spectrum we find the function of "sound reference" (semanticity, referentiality, lexical meaning) where arbitrary sound patterns are used to convey symbolic meaning. At the other end we find "sound emotion," where rather particular sound patterns (either culture-specific or species-specific) are used to convey emotional meaning.[3] According to this view, music and language differ mainly in their emphasis rather than in their fundamental nature, such that language emphasizes sound reference while downplaying its sound emotion aspect (although it certainly makes use of sound emotion), whereas music's acoustic mode emphasizes sound emotion while downplaying its referential aspect (although it certainly makes use of referentiality). Language and music are essentially reciprocal specializations of a dual-natured precursor that used both sound emotion and sound reference in creating communication sounds. However, along with this reciprocal specialization, various functions appear in the middle of the spectrum in figure 16.1 that bring these two specialized capacities together. From the music pole comes music's vehicle mode of action, in which language's referentiality and music's sound emotion function come together in a complex union of reenactment rituals, musical symbolism, musical narration, acoustic depiction, and the like. From the language pole comes a whole slew of features involved in heightened speech, *sprechstimme*, rapping, *recitativo*, poetic meter, and the abundant pragmatic uses of speech melody and rhythm to convey linguistic and paralinguistic meaning. Thus, the task of the musilanguage model is to describe a system containing both rudimentary referential and sound emotion properties such that it might be a reasonable precursor for the evolution of both music and language, and such that divergence from this precursor stage can be seen as an intensification of emphasis rather than the creation of new worlds.

The Musilanguage Stage

The present section attempts to characterize the necessary properties of the musilanguage stage, and later sections present a description of the origins of this stage as well as the divergence process that led to the formation of music and language. As will be seen shortly, development of these ideas was inspired quite a bit by phonological theory in linguistics, which has (surprisingly) played an even smaller a role in theories of language origin than it has in theories of music origin. The idea that speech and music are systems of expressively intoned sound is well accepted. But what is often ignored is the extent to which intonational concerns for melody, rhythm, and phrasing in speech strongly parallel those in music, not just in a metaphorical sense but in a mechanistic sense.

Several properties of the musilanguage stage contribute to the shared ancestral features of music and language. To facilitate discussion of a complex topic, a summary of the argument will guide the reader. I contend that at least three essential features of a musilanguage device are necessary for it to qualify as a precursor and scaffold for both language and music.

1. *Lexical tone*: use of pitch to convey semantic meaning. This involves creation of a tonal system based on level tones (discrete pitch levels).

2. *Combinatorial formation of small phrases*: generation of phrases by the combinatorial arrangement of unitary lexical-tonal elements. These phrases are melodorhythmic as well as semantic structures. One source of phrase melody is the sequential organization of the pitches contributed by the elemental units. A second one consists of global melodic formulas.

3. *Expressive phrasing principles*: use of local and global modulatory devices to add expressive emphasis and emotive meaning to simple phrases. Four general mechanisms of phrasing are envisioned that modify the acoustic features of the phrase to create basic intonational phrases.

Evolutionarily, this is seen as emerging through a two-step process in figure 16.3, proceeding from a primary stage of single lexical-tonal units

Three Sources of
<u>Phrase Melody:</u>

<u>First Musilanguage Stage:</u>

1) LEXICAL TONE
Use of pitch to convey semantic meaning
Discreteness of units: pitched vocalizations; level tones
Broad semantic meaning

<u>Second Musilanguage Stage:</u>

2) COMBINATORIAL PHRASE FORMATION
Simple combinations of lexical-tonal elements *sum of local pitch contours*
¤ melodic and rhythmic units
Two levels of meaning:
¤ local: relations between unitary lexical elements
¤ global: phrase-level (emotive) meanings *global melodic formulas*

3) EXPRESSIVE PHRASING
Four levels of phrasing: *expressive modulation*
¤ global/graded: global sentic modulation
¤ global/categorical: contour/meaning associations
¤ local/graded: local sentic modulation (prosody)
¤ local/categorical: prominence (accent/stress)

Figure 16.3
Summary of the properties of the musilanguage stage. The model highlights three important properties of the putative musilanguage precursor. Three general properties are thought to provide an adequate description of the precursor of both music and language, and emerge in the form of two distinct stages. The first musilanguage stage is a unitary lexical-tonal system. This involves a system of discrete and pitched vocalizations that are functionally referential in a very broad sense. The second musilanguage stage simultaneously introduces phrase formation and phrasing. Phrase formation is based on simple combinatorial principles involving lexical-tonal elements introduced during the first musilanguage stage. Four mechanisms of phrasing are also introduced that modulate the acoustic properties of these combinatorially generated phrases, as described in the text. Phrase melody is thought to receive three independent but related contributions: the sum of lexical-tonal elements, global melodic contours, and expressive modulation.

(first musilanguage stage) to a later stage of phrase formation based jointly on combinatorial syntax and expressive-phrasing principles (second musilanguage stage). These three overall properties are thought to make independent but related contributions to the global melody of a musilinguistic phrase, as shown on the right side of figure 16.3.

Lexical Tone

This refers to the use of pitch in speech to convey semantic (lexical) meaning. Languages that make extensive use of lexical tone as a suprasegmental device are called tone or tonal languages. As they tend to be viewed as oddities by linguists, theories of language origin tend to ignore the fact that not merely a handful of exotic languages fall into this category, but that a majority of the world's languages are tonal (Fromkin 1978). The most parsimonious hypothesis is that language evolved as a tonal system from its inception, and that the evolutionary emergence of nontonal languages (intonation languages) occurred due to loss of lexical tone. In other words, this hypothesis states that tonality is the ancestral state of language. Intermediate cases exist, called pitch-accent languages, exemplified by Japanese, Swedish, and Serbo-Croatian, in which some limited use of contrastive tone is employed in the presence of intonation. Such limited uses of tone might represent either remnants of an earlier tonal stage, or, as is the case for Swedish and Norwegian, secondary acquisition of tonal properties from a nontonal precursor. As tone can be both acquired by and lost from languages, the goal here is not to describe the history of individual languages, but to describe the evolutionary history of language as a whole. I think that there are good evolutionary reasons for believing that tonality was the ancestral state of language, but this will have to be explored elsewhere.[4] The major point is that the notion of lexical tone implies that pitch can and does play an essential role in language, not just as a prosodic or paralinguistic device, but as a semantic device.

The single biggest complication in viewing lexical tone as a musilinguistic feature rather than a purely linguistic feature is the problem of level tones or pitch levels. Whereas all musical systems consist of sets of discrete pitches, intonation languages such as English appear on first view to make no such use of discrete pitch levels, but instead seem merely to be waves of sound punctuated by prosodic accents. It is here that my thinking is greatly indebted to autosegmental theories in phonology (Goldsmith 1976, 1990; Pierrehumbert 1980/1987; Ladd 1996). Historically, there has been a long-standing debate in phonology between a so-called levels perspective and a so-called configurations or contours perspective; that is, whether intonational events should be best thought of in terms of sequential movements between discrete pitch levels, or in terms of the pitch movements themselves irrespective of any notion of level tones. In the former view, pitch contours are merely transitions or interpolations between discrete pitch levels, whereas in the latter view they are the phonological events of interest. Many important phonological issues hinge on this levels-versus-configurations debate. Autosegmental theory was hailed as a resolution to this controversy (Ladd 1996). It supports the levels view by saying that phonological events should be modeled as sequential movements between discrete pitch levels, often only two levels, High and Low, and that all movements between them should be reduced to the status of transitions, rather than primary phonological events of importance (Goldsmith 1976). Thus, the notion of level tones is central to autosegmental theory, but, of importance, this applies as much to intonation languages as it does to tonal languages. Autosegmental theory confers onto level tones a status of general importance in all spoken language. In addition, it imposes an explicitly localist view on phonology, regarding all spoken utterances as series of steps from one level tone to the next. Two additional tonal features, downstep and boundary tones, are sufficient to confer onto utterances the overall wave-like properties that configurations supporters focus on (Pierrehumbert 1980/1987; Ladd 1996).

Autosegmental models have been applied to many languages, tonal and intonation alike (see Goldsmith 1995; Ladd 1996), and cognitive experiments have been highly supportive of the autosegmental interpretation. Ladd (1996) presents a general model of pitch-range effects from the autosegmental perspective that is of general relevance to the musilanguage model. Ladd contrasts two different ways of thinking about pitch in speech: an initializing approach in which phonological pitches are defined with reference to neighboring pitches (e.g., pitch Y is three semitones higher than proceeding pitch Z, and two semitones lower than preceding pitch X), and a normalizing approach in which such pitches are described in normalized terms with reference to their position on a scale describing a speaker's total pitch range (e.g., pitch Y is 80% of the speaker's highest pitch; alternatively, pitch X is 1.75-fold higher than the lowest frequency in the speaker's pitch range). Ladd supports the normalizing model, and it makes the most sense in terms of the current model.

Within the context of the autosegmental theory's focus on level targets, the normalizing approach to pitch predicts that scaling of these level targets should be systematic between speakers, and this is exactly what several studies showed (Thorsen 1980, 1981; Liberman and Pierrehumbert 1984; Ladd and Terken 1995). In other words, when multiple speakers are asked to read multiple sentences in a given language, and the absolute frequencies are normalized with respect to the speakers' pitch-range, an extremely high correlation (around .9) is found between target values of one speaker and those of another. The utterances are scaled. The scale may change as a function of pitch level (raising or lowering one's voice) but does not vary among speakers having different vocal ranges. The general implication of these findings for the musilanguage model are striking. They hold that speech, like music, is based on scales consisting of discrete pitch levels. The major difference between speech and music in this regard is that these scales change quite a bit during speech (e.g., when pitch level changes) and thus so do the level tones themselves. But this does not negate the basic observation that the scaling of pitch is used in speech, as predicted by the normalizing-autosegmental approach to pitch range.

Another important point that has bearing on the use of tone in speech is the observation of categorical perception of tone. House (1990) presented his experiments with Swedish speakers and reviewed the literature with regard to Chinese lexical tone, German categories of intonational meaning, and English pitch accent, and concluded that "results from perception experiments in four different languages support the concept of linguistic categories (both lexical and semantic) being perceived in terms of tonal levels during maximum spectral change after the CV [consonant-vowel] boundary and as tonal movement during relative spectral stability. The synchronization of tonal movement with vowel onset seems to be important for the perception of linguistically relevant tonal categories" (p. 81). Thus for both intonation languages and tone languages, cognitive experiments show that people tend to perceive level tones in a more or less categorical fashion, in support of autosegmental models of intonation and lexical tone.

What are the implications of these important findings for the musilanguage model? Three basic implications bear mentioning. First, the production and perception of *pitched* vocalizations is a necessary characteristic of such a system, in contrast to vocalizations based purely on portamentos (glides, slides, etc.). As most primate vocalizations systems rely heavily on unpitched grunts and pants (e.g., chimpanzee pant-hoots, vervet monkey alarm calls) or on high-contoured pitch glides (gibbon song), the musilanguage theory posits that a pitched vocalization system involving at least two pitch states would have had to evolve at some point in the hominid line. This theory does not demand evolution of new articulatory capacities to form novel types of segmental phonemes but simply the cognitive capacity to use level tones in a meaningful fashion. Nor does this argument have any bearing on the types of transitions that occur between level tones; they are just as likely to be pitch glides as pitch jumps. All that is important is that some notion of level tones be involved.

Second, the idea of lexical tone, as seen from the autosegmental perspective, suggests that level tones are just as important for intonation languages as they are for tone languages. Therefore, discrete pitch levels and pitch-scaling mechanisms are not merely features of tone languages and music but are important features of intonation languages as well. Speech, like music, is based on discrete pitch levels that themselves are scaled, although variably so. This is supported by experiments showing that normalizing approaches explain pitch-range effects better than do initializing approaches as well as by studies demonstrating the categorical perception of tone in both intonation languages and tone languages.

Third, any evolutionary expansion of this system to generate phrases will follow, at least to an important extent, localist rules whereby strings are assembled in a sequential, stepwise fashion (this is described in more detail below). The insight from autosegmental theory for the musilanguage model is that sequences of level tones can be the basis for semantic strings. The fact that intonation languages dissociate such strings of level tones from semantic strings emphasizes the earlier point that language's meaning level has no obligatory relationship to its phonological level or even to the acoustic modality. Intonation languages, like gesture languages, highlight the primary importance of creating semantic meaning from meaningless components, whatever these components may be. However, the evolutionary hypothesis here is that language began as a tonal system, and this seems to be borne out, at least in part, by the robust presence of lexical tone in the world's languages.

Finally, a natural question that emerges is, how can I argue that a system of lexical tone could be a precursor for music? Isn't music based on meaningless pitches rather than meaningful lexical units? This is a question that is central to the issue of musical semantics. First of all, I mentioned that divergence from the musilanguage stage would lead to differences in emphasis between music and language. So it is only natural to think that music would deemphasize its lexical tonal aspect during this divergence process. Yet at the same time, two other points have a bearing on this issue. The first is to emphasize that lexical words can have, and often do have, a very broad range of meanings, where semantic interpretation is highly dependent on the context of not only the sentence but the entire discourse arrangement. Thus, words have great semantic elasticity (Swain 1997), and this is seen in abundance during the development of speech in children, where lexical words start off having extremely broad meanings, and acquire precise meanings only as the lexicon and syntactic system expand during later stages of development. The second idea is that music has many devices available to it to give it semanticity. This was discussed above with reference to music's vehicle mode of action, especially in relation to the use of music for symbolization and narration (see note 2).

One example of this is the leitmotif in Western opera, where particular musical motifs become semantic tags for characters, objects, or concepts. Another example consists of drummed and whistled languages (Umiker 1974). There is no question that the semantic system of the musilanguage stage would have to have been very broad for lexical tone to qualify as a shared ancestral feature of music and language. However, ". . . a passage of music could have a semantic range that is essentially the same as that of any word in a language, only much broader in its scope; sharing the same kind of elasticity but of much greater degree than is typical in language" (Swain 1997:55). In sum, I believe that the notion of lexical tone, with its underlying level tones and semantically meaningful pitch movements, satisfies the criterion for being a joint feature of language and music, and a scaffold on which both systems could have developed. This first musilanguage stage would have been a system of unitary lexical-tonal elements which could have been combined to form phrases.

Combinatorial Phrase Formation

Given the establishment of a lexical tone-based vocalization system, we can envision the next evolutionary step in the musilanguage system’s development whereby sequences of lexical-tonal units are strung together to make simple, unordered phrases having higher-order meanings. The semantic meaning of such phrases has both compound and global sources. The compound sources are derived from the relational juxtaposition of the individual semantic units being combined. A global level of meaning, due to the overall melodic contour of the phrase, is a second important semantic feature of a phrase-based system not possible in a single-unit system, such as the first musilanguage stage. These phrase-level melodies correspond to categorical formulas for conveying emotive and/or pragmatic meaning (see Richman, this volume). In the domain of speech, they include such discrete phonological formulas as question intonations and surprise intonations. Thus, phrase-based systems provide a dual advantage over single-unit systems in that they have two levels of meaning: compound—meaningful relations between the individual units, and global—categorical formulas characterizing the phrase as a whole. Such combinatorial phrases have not only a melodic structure but a rhythmic structure as well, and the rhythmic patterns of such phrases are derivable, at least in large part, from the temporal arrangement of elemental units.

I maintain that whereas the basic ingredients of hierarchical organization are present in such a system, this second musilanguage stage has neither a sense of ordering nor a strong sense of hierarchical grouping. The one exception to this, described below, is the notion of prominence. In general, hierarchical organization would have emerged in a modality-specific fashion after divergence from the musilanguage stage, leading to the creation of the specific grammars of language and music. Therefore, one important implication of this model is that *the general capacity for combinatoriality preceded the evolution of modality-specific syntaxes* As such, this model shares features with Bickerton’s (1995) protolanguage model. The musilanguage stage should have had neither the propositional syntax of language nor the blending syntax of music, but should have merely been a system of combinatorial relations between basic elements in which an additional, global level of meaning was superimposed on the relational level of meaning. However, despite this absence of a complex syntax system, this second stage is a richer and more flexible communication system than a single-unit system in that it provides at least two levels of meaning from a single phrase. Thus, in one sense, phrases are simply the sum of their parts (localist features), but in another sense they are something more than the sum of their parts (globalist features).

The biggest complication of this model lies in trying to tie together combinatorial phrase formation with autosegmental ideas of level tones in speech. The case of music is far simpler. Virtually all of the world's musical systems are based on sets of discrete pitches, subsets of which are used to generate motifs and melodies. To what extent can we think of speech as being a melodic generative system in the same way? Pierrehumbert and Hirschberg (1990) proposed a localist, compositional approach to the production of phonological phrases that is based on the simple bitonal features of autosegmental models. However, such models have no explicit requirement that the High and Low level-tones correspond to anything like the discrete absolute-frequency (F_0) levels that go into formation of musical scales. Yet, my own argument is critically dependent on this. This was mentioned above in relation to lexical tone. I think that the resolution to the problem is to reconsider Ladd's (1996) normalizing approach to pitch features and say that whether people are actually aware of it or not, they tend to use pitch in a scaled fashion in producing speech utterances. In fact, I think the situation is no different in musical generative systems. People create melodies or songs using implicit cognitive rules based on the discreteness of pitch, which is dependent on the categorical perception of pitch (Lerdahl 1988). Phonological evidence suggests that people do something quite similar when speaking, thus supporting the basic combinatorial pitch arrangement in speech. So the general conclusion here is that speaking is not only pitched but scaled, and that people obey scaling principles in generating speech utterances. By this analysis, speech melody is no longer a metaphor, but a mechanistic parallel to musical melody, itself based on scaled pitches.

Expressive Phrasing

Cognitive musicology has placed such a premium on exploiting the grammar metaphor in music that it has all but ignored many important parallels that occur at the level of intonational phrasing. Generative theories of music have been rightly criticized for their failure to address these expressive properties, such as tempo, dynamics, rhythmic modulation, and the like. It is not sufficient for musical phrases to have hierarchical melodic and rhythmic structure; they must also function as intonational phrases for the expression of emotion and emphasis. But the most important point to emerge is that expressive phrasing is so general that it is wrong to dichotomize its forms in speech and music. Phonologists describing speech phrasing and musicologists describing musical phrasing often talk about exactly the same processes, but with two different sets of terms. Therefore it is important to subsume these phrasing mechanisms into a unified set of concepts and terms (figure 16.4) that are rooted in biological notions of common evolutionary ancestry.

Before talking about these mechanisms, I would like to introduce one concept that has general relevance to this topic: sentic modulation. The term “sentic” I borrow from Manfred Clynes (1977); however, I do not use it in exactly the same sense that Clynes did. I use it in a more limited sense, as expressed in Clynes’ equivalence principle: “*A sentic state may be expressed by any of a number of different output modalities* . . . gestures, tone of voice, facial expression, a dance step, musical phrase, etc.” (p. 18, emphasis in original). My take on Clynes’ equivalence principle is to say that the sentic system is a general modulatory system involved in conveying and perceiving the *intensity* of emotive expression along a continuous scale. It expresses intensity by means of three graded spectra: tempo modulation (slow-fast spectrum), amplitude modulation (soft-loud spectrum), and register selection (low-pitched-high-pitched spectrum). This system appears to be invariant across modalities of expression in humans, such as speech, music, and gesture, on which Clynes’ equivalency is based. It also appears to function in a similar way in emotive behavior in nonhuman animals (Morton 1977, 1994),

Fig. 16.4

Figure 16.4
Four mechanisms of expressive phrasing are described along two dimensions, acting first either at the global level or the local level of the phrase, and second in either a graded manner (local and global sentic modulation) or in a more discrete, categorical manner (contour-meaning associations and prominence effects). See text for details.

suggesting that the sentic system might be one feature of musical processing that has homologues in vertebrate expressive behavior generally. The universality of this system for human emotional expression can be demonstrated by pointing out that in speech, gesture, and music, the same sentic profile occurs to express a given emotional intensity state, regardless of the modality of expression. For example, happy music and happy speech are both characterized by fast tempos, large-amplitude sounds, and high registers; sad music and sad speech are characterized by the opposite sentic spectrum. Looking at gesture instead of vocalization, one sees that happy movements are characterized by fast tempos, large amplitudes (broad gestures), and high positioning (the equivalent of high register), with sad gesturing showing exactly the opposite spectrum. In all cases, the level of sentic modulation reflects the intensity level of emotional expression, thus highlighting the gradient nature of the sentic system. Happy movements are fast, but ecstatic movements are ballistic; sad movements are slow, but depression is immobilizing. Again, much evidence suggests that sentic modulation is not merely cross-modal, but also cross-cultural and cross-species. Sentic factors are an excellent place to look for universal expressive features in music, speech, and gesture.

Four general mechanisms of expressive phrasing are used in speech and music. As seen in figure 16.4, they are divided along two dimensions, acting either at the local or global levels of the phrase, and acting in either a graded or categorical fashion with respect to the acoustic parameters being modulated.

Global Level

We can think about two phrasing mechanisms acting at the global level (figure 16.4, left side): global sentic modulation and contour-meaning associations. Global sentic modulation involves expressive devices that affect the intensity level of the whole phrase with regard to overall tempo, amplitude, and register. These effects occur along a continuous spectrum such that the level of sentic modulation correlates with the intensity of emotional expression. As mentioned, global sentic effects have the same emotional meaning in music and speech, and the sentic profile for a particular emotional state in music and speech is exactly the same.

The second factor of global expressive phrasing involves all categorical contour-meaning associations that relate phrase melody to particular meanings. Unlike global sentic modulation, contour-meaning associations work in categorical fashion, with each melody having a more or less specific meaning (see Richman, this volume). Things such as question intonations, surprise intonations, and call intonations are universal melodies that convey pragmatic features of discourse. Similarly, in Western music, "question phrases" (ascending contours) convey a feeling of tension and uncertainty, whereas "answer phrases" (descending contours) convey a feeling of resolution of that uncertainty. Interestingly, in both speech and music, ascending contours convey uncertainty and uneasiness, and descending contours certainty and stability, providing further evidence that these phrasing mechanisms arose from a joint precursor. As mentioned earlier, compositional approaches to speech intonation (Pierrehumbert and Hirschberg 1990) tend to reduce global phrase-level formulas to local-level sequential tone changes. Be that as it may, such formulas tend to operate in a global, categorical fashion.

Local Level

Two phrasing processes act at the local level (figure 16.4, right side): local sentic modulation (prosody) and prominence. Prosody encompasses our most basic idea about intonation, referring to the local risings and fallings, quickenings and slowings, and loudenings and softenings that are involved in expressively conveying our meanings in a pragmatic sense. To my mind, prosody is best represented as a series of sentic rules acting at the local level. These rules are in principle similar to those acting at the global level except that they act locally, involving small modulations in tempo (accelerando, ritardando), pitch (ascent, descent), volume (crescendo, diminuendo, sforzando), and length (ritenuto) at the level of the individual element or group of elements. As with global sentic modulation, local modulation occurs along a continuous intensity gradient, and this gradient effect is certainly one of the most important characteristics of speech intonation and musical phrasing. This level of phrasing is one feature that distinguishes one speaker from another or one musician from another.

The second local phrasing mechanism involves use of accent or stress as prominence devices to convey emphasis or focus in either speech or musical phrases. A phrase usually has a single point of emphasis, thus making prominence a categorical signal acting at the local level. There are several ways of effecting prominence: a rise in pitch, an increase in amplitude, an increase in duration, or some combination thereof. Local sentic modulation (prosody) and prominence interact in such a way that the part of the phrase that precedes the accent often demonstrates a continuous build-up, whereas the part that follows it shows a continuous fall-off. In both music and speech, prosody is used in the service of prominence by allowing phrases to be elaborated in a smooth rising-and-falling fashion, rather than in a punctuated manner.

These four phrasing mechanisms affect the ability of speakers and musicians to convey emphasis, emotional state, and emotional meaning.

Whether in speech or music, they modulate the same basic set of acoustic parameters, making interdependent contributions to the process of phrasing.

Summary

To summarize this section, I propose an evolutionary progression from a simple system involving a repertoire of unitary lexical-tonal elements (first musilanguage stage) to a less simple system based on combinatorial arrangements of these lexical-tonal (and rhythmic) elements (second musilanguage stage). The latter obtains its meaning not just from the juxtaposition of the unitary lexical elements but from the use of global phrase-level melodies. It is at the same time a phrasing system based on local and global forms of sentic modulation as well as on prominence effects. One offshoot of this analysis is that phrase melody has three important but distinct sources (figure 16.3): the sum of the local pitch contours from the lexical-tonal elements; phrase-level, meaningful melodies; and intonational modulation through expressive phrasing mechanisms. An important evolutionary point is that combinatorial syntax is seen to precede modality-specific grammars. This system is, to a first approximation, a reasonable precursor for the evolution of both music and language out of which both could have emerged while retaining the many important properties they share.

Before closing this section, it would be useful to return to the question of generativity and hierarchical organization. I stated at the beginning of the chapter that generativity is an analogous feature of language and music, not a shared ancestral feature. Music's and language's generativity are based on completely different syntactic principles whose only common denominators are discreteness and combinatoriality. At the same time, it is not difficult to imagine hierarchical organization evolving out of the musilanguage precursor stage, thereafter becoming exploited by modality-specific systems. All that is necessary is for some type of either grouping or segregation of elements (or both) to occur to differentiate different elements within the phrase. This could occur at the level of pitch (auditory streaming effects), rhythm (pulse relationships), amplitude (prominence effects), and so on. The point is that the musilanguage device, based on discreteness, combinatoriality, and intonation, provides all the necessary ingredients for hierarchical organization in what will eventually become two very different grammatical systems. So the actual forms of hierarchical organization in music and language are best thought of as resulting from parallelism rather than from common origins, again with the note that the shared ancestral features of the musilanguage stage provide fertile ground for evolution of hierarchical organization once the divergence process starts to take off. The only hierarchical function that seems to be a necessary part of the musilanguage stage is prominence. Acoustically, prominence can be effected by a diversity of mechanisms, including pitch, length, and strength.

Precursors

Given this analysis of the musilanguage stage as a joint precursor of music and language, two major questions remain: what are the origins of the musilanguage stage? and what is the process by which the divergence occurred to make music and language distinct, sometimes dichotomous, functions along the spectrum described in figure 16.1?

Regarding the first question, one hint comes from a very interesting and well-described class of primate vocalizations, which I call *referential emotive vocalizations*. A referential emotive vocalization (REV) is a type of call (not song) that serves as an on-line, emotive response to some object in the environment, but that also has the property of semantic specificity for the class of object being responded to. Thus, each call-type signifies a given object. From the standpoint of nearby conspecifics, REVs serve an important communicative function for the social group, as the meaning of each call is known to all members of the species, thereby encouraging appropriate behavioral responses. For the purposes of this discussion, the most salient feature of a REV is its dual acoustic nature: a given sound pattern has both emotive meaning and referential meaning, a property shared with the musilanguage stage that I proposed.

The best-described referential emotive system is the alarm call system of the East African vervet monkey, which has a repertoire of at least three acoustically distinguishable calls (Struhsaker 1967). In fact primates and birds have a large number of such functionally referential calling systems that have a similar level of semanticity to that of vervet alarm calls (see table 3.1 of Marler, this volume; Hauser, this volume; Marler, Evans, and Hauser 1992). Acoustically, vervet calls are short grunts that are specific for the predator eliciting the alarm. The best-characterized calls are the eagle, snake, and leopard calls. That vervet monkeys know the meaning of the calls is shown by audioplayback experiments in which the animals engage in appropriate escape behaviors to the different calls, running up into trees on hearing the leopard call, and looking to the sky or running into bushes on hearing the eagle call (Seyfarth, Cheney, and Marler 1980a, b). At the semantic level, REVs show the same type of broad semantic meaning that is suggested for the musilanguage device.

I propose that the precursor of the musilanguage stage was a type of REV. It is not important that this be an alarm call system per se, but merely a system with its characteristic dual acoustic nature and broad semantic meaning. The most important feature that would have been required to move from a vervet-type REV to the first musilanguage stage would have been the meaningful use of discrete pitch levels, in contrast to the unpitched grunts of many primate calls. Although such a system has not been described, the vervet alarm call system holds out as an important model for how it might operate, providing clues as to how the musilanguage stage may have evolved.

Divergence

The second question was, by what process did the divergence from the musilanguage stage occur to make music and language distinct though related functions? How did language become "language" and music "music" starting from the hypothesized musilinguistic ancestor? This question relates most directly to the origins of language and music as they occur in their current forms. My goal is not to rehash the extensive series of functional theories that have been proposed to account for the origins of human language (reviewed in Wind et al. 1992; Lewin 1993; Beaken 1996), but to see how the current proposal of a joint musilanguage stage affects such theories. Let us look again at the functional spectrum presented in figure 16.1. As stated, music and language sit at opposite ends of a spectrum, with each one emphasizing a particular type of interpretation of communicative sound patterns. The two evolved as reciprocal elaborations of a dual-natured referential emotive system, again suggesting that they differ more in emphasis than in kind.

In thinking about the divergence process, it is useful once again to return to the distinction among shared ancestral, analogous, and distinct features of music and language. By definition, the first type of feature appeared before the divergence process and the second two after it. Divergence can therefore be characterized as the process by which the analogous and distinct features of music and language evolved. However, this probably came about two different ways. Analogous features most likely represent specializations emerging out of the shared ancestral features of the musilanguage stage. They are differentiation events. Distinct features, such as music's isometric rhythms and language's propositional syntax, are not. Instead they represent modality-specific (and human-specific) novelties of these two functions. Let us now consider these features.

Looking first to language, we see that this system not only develops an explosively large lexicon (some 100,000 words in adult humans), but a semantic system containing greatly specified meanings by comparison with a primate REV or the musilanguage system. At the level of grammar, language develops a kind of propositional syntax that specifies temporal and behavioral relationships between subjects and objects in a phrase. Because it makes reference to personal experience, this syntax system can be the basis for determinations of truth and falsity. Structurally, it involves not only simple hierarchical organization but recursiveness as well. Perhaps the point of greatest distinction from music is language’s liberation from the acoustic modality altogether, leading to amodal conceptualization, off-line thinking, and human reason.

Looking to music, divergence from the musilanguage stage leads initially to the formation of its acoustic mode. The acoustic range and pitch repertoire become greatly expanded over anything seen in the musilanguage precursor or in spoken language, extending to more than eight octaves, each octave being divisible into at least a dozen differentiable pitches. At the level of grammar, music acquires a complex and hierarchical syntax system based on pitch patterning and multipart blending, leading to the creation of diverse motivic types, many forms of polyphony, and complex timbral blends. In addition to this pitch blending property, we see the emergence of many categorical formulas for expressing particular emotional states, leading to the various forms of sound emotion that are used in creating coherent and emotively meaningful musical phrases. Finally, at the rhythmic level, music acquires the distinct feature of isometric time keeping, so much a hallmark in Western culture. This metric-pulse function is based on a human-specific capacity to both keep time and to entrain oneself rhythmically to an external beat. This permits rhythmical hierarchies in both the horizontal and vertical dimensions of musical structure, including such things as heterometers and polyrhythms.

Evolutionary divergence results in significant differences between music and language at the highest levels. The last thing to explain is how these two systems came together to create yet newer functions. For this, it is important to distinguish between the shared properties and interactive functions. Shared properties of music and language are posited by the musilanguage model to be either shared ancestral or analogous functions. Interactive functions are areas in which music and language come together to create novel functions that strongly involve both systems. This was demonstrated on the spectrum presented in figure 16.1. It includes principally all those functions that I call the vehicle mode of music operation, not to mention the use of meter in poetry and the many exaggerated uses of intonation to convey information, attitude, and emotion. The major point is that interactive functions develop through a coevolutionary process that reflects the evolutions of both the linguistic and musical systems. For this reason, we expect interactive functions, such as verbal song, to evolve through a series of stages that reflect the evolution of the two systems contributing to these novel functions.

Summary

The full musilanguage model can now be presented. It posits that music and language evolved as two specializations from a common ancestor, such that a series shared ancestral features evolved before either analogous or distinct features. This model is distinguished from those holding that music evolved from a dedicated linguistic capacity (music outgrowth model) or that language developed from a dedicated musical capacity (language outgrowth model). It argues instead that shared ancestral features of music and language should be thought of as musilinguistic rather than either musical or linguistic. The model's principal contribution to the study of language evolution is to provide a new chronology for the development of language's structural features: language evolved out of a sophisticated referential emotive system; phonological syntax preceded propositional syntax; tone languages preceded intonation languages; speech could have evolved early, due to its exploitation of lexical tone instead of enlarged segmental inventories; lexical tone, combinatorial syntax, and expressive intonation were ancestral features of language that were shared with music; broad semantic meaning preceded precise semantic meaning; and language's acoustic modality preceded its representational state of amodality.

The model begins with a referential emotive system (figure 16.5) that in its most basic form provides for the dual acoustic nature of the musilanguage system: sound as emotive meaning and sound as referential meaning. This by itself establishes the functional spectrum that will later define music and language as two separate specializations. From this we see the development of the musilanguage stage, which is thought to have occurred in two steps. The first step was the use of level tones (discrete pitches) and pitch contours for referential communication. The second step was the development of meaningful phrases, generated through combinatorial rules for joining discrete elements into phrases, these phrases being subject to four levels of modulation: local sentic rules for expressive modulation; global sentic rules for the overall level (intensity) of expression; local categorical rules for prominence; and global categorical formulas for generating phrase-level contour-meaning associations. These devices make independent but related contributions to the overall acoustic properties of the phrase. Semantically, the musilanguage device is a sophisticated referential emotive communication system that generates meaning at two levels: first, from the relational juxtaposition of unitary elements (local level), and second, from overall contour-meaning associations (global level).

The next step in this evolution is the simultaneous occurrence of divergence and interaction, with continued retention of the shared ancestral

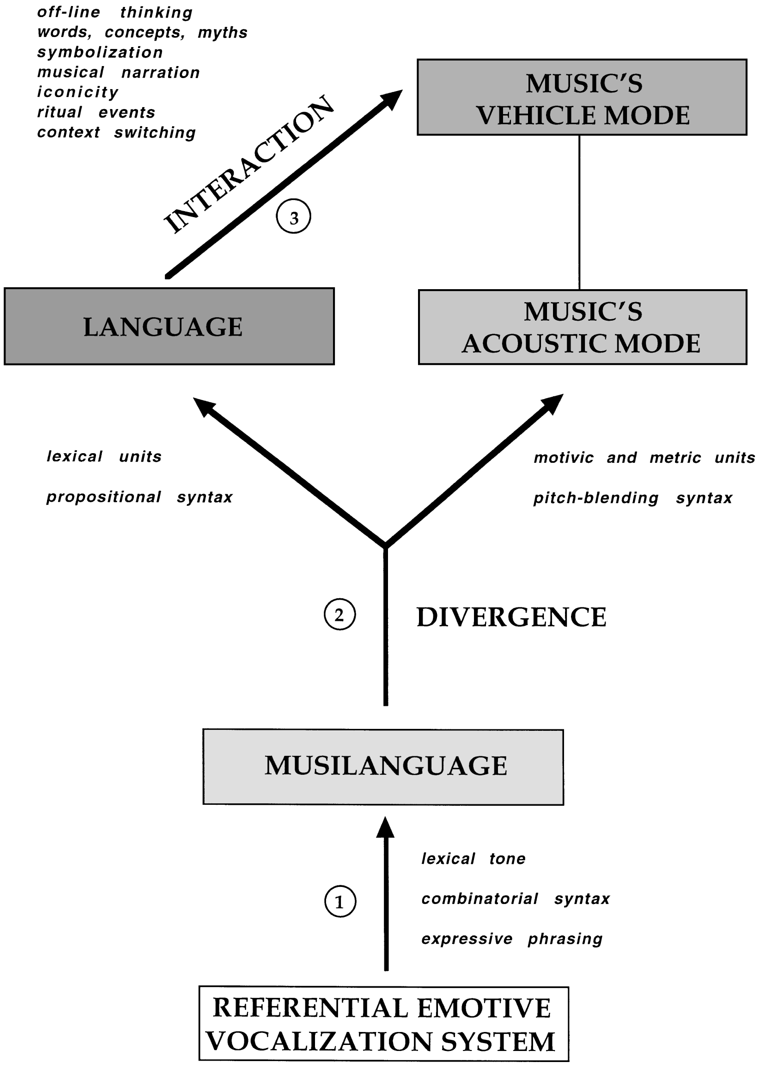

Figure 16.5
The full musilanguage model begins with a hominid referential emotive vocalization system, which provides for the dual acoustic nature of the musilanguage stage: sound as referential meaning and sound as emotive meaning. Next, the musilanguage stage is thought to evolve by a two-step process, beginning first with a unitary lexical-tonal system, followed by a phrase system involving both combinatorial syntax and expressive phrasing properties. This musilanguage stage provides for the shared ancestral features of music and language. The next step is divergence from the musilanguage stage, leading eventually to the mature linguistic system and music's acoustic mode. This occurs through reciprocal elaboration of either sound as referential meaning (language) or sound as emotive meaning (music's acoustic mode). This involves not only different fundamental units at the phonological level but different interpretations of these units at the meaning level. An important aspect of the divergence process is the formation of different syntax types: propositional syntax in the case of language, and blending syntax in the case of music. The final step is development of interactive properties by a coevolutionary process. This leads to, among other functions, music's vehicle mode of action, which involves such things as verbal song, iconic representation, and musical narration (see footnote 2 for details).

features. Divergence occurs due to the reciprocal elaboration of either sound as referential meaning or sound as emotive meaning, ultimately making language and music different in emphasis rather than in kind. This is accompanied by an important divergence of syntax types: language's propositional syntax is based on relationships between actors and those acted upon; music's blending syntax is based on pitch blending and pitch patterning leading to complex sound-emotion relationships. This establishes language's symbolic capacity for representation and communication and music's acoustic mode (with its sound-emotion system and broad semantics). Finally, simultaneous with the divergence process is the formation of interactive functions, exemplified by verbal song and all the other vehicle functions of music. In other words, divergence is accompanied by rebinding of music and language in the form of novel functions that evolve parallel to their separation. The emergence of these interactive functions reflects coevolution of the underlying linguistic and musical systems. Thus, we can imagine verbal song as evolving through a series of stages that parallel biological developments in both systems.

What of functional evolutionary concepts? I do not think anyone would deny that both music and language are highly multifunctional. However evolutionary models are adaptationist interpretations of how traits evolve, and tend to focus monolithically on a single adaptive function and a single selection mechanism for a given trait. So far, the monolithic approach to language has failed miserably, and I doubt that it will work for music either. But in addition, and more controversially, I sincerely doubt that functionalist concepts of music origins based exclusively on individual selection processes will, in the end, bear fruit. There is just too much about music making that reveals an essential role in group function to ignore the issue of multilevel selection (Sober and Wilson 1998). Nobody questions that music is done in groups, but Miller (this volume) seriously questions whether it is done *for* groups. Half a century of ethnomusicological research suggests that a principal function, if not *the* principal function, of music making is to promote group cooperation, coordination, and cohesion (Merriam 1964; Lomax 1968; Hood 1971). Music making has all the hallmarks of a group adaptation and functions as a device for promoting group identity, coordination, action, cognition, and emotional expression. Ethnomusicological research cannot simply be brushed aside in making adaptationist models. Contrary to strong sexual selection models, musical activity in tribal cultures involves active participation by the entire group, that is, both sexes and people of all ages. Such cultures make no distinction between musicians and nonmusicians. Where sex or age segregation is found at the level of performance style, it is usually a reflection of specialization at the level of the work group (Lomax 1968), and this is described by the universal ethnomusicological principal of functionality or context specificity in musical performance. Music making is done for the group, and the contexts of musical performance, the contents of musical works, and the performance ensembles of musical genres overwhelmingly reflect a role in group function. The straightforward evolutionary implication is that human musical capacity evolved because groups of musical hominids outsurvived groups of nonmusical hominids due to a host of factors related to group-level cooperation and coordination.

Finally, as a tie-in to our discussion of the musilanguage model and the divergence process leading to music’s outgrowth from the musilanguage precursor, music has two distinct design features that reflect an intrinsic role in group cooperation. These two features account for a large part of what music is at the structural level: pitch blending and isometric rhythms. Whereas speech proceeds obligatorily by an alternation of parts, music is highly effective at promoting simultaneity of different parts through its intrinsic capacity for pitch blending; music’s vertical dimension must be seen as a design feature for promoting cooperative group performance and interpersonal harmonization. In addition, musical meter is perhaps the quintessential device for group coordination, one which functions to promote interpersonal entrainment, cooperative movement, and teamwork. Pitch blending and metric rhythms are central to any evolutionary account of the melodic and rhythmic dimensions of music. *Theories of individual selection must explain how these essentially group-cooperative musical devices evolved in the service of within-group competition*. I doubt that such models will be able to account for them, and I suggest instead that multilevel selection models involving group selection (Sober and Wilson 1998) and/or cultural group selection (Boyd and Richerson 1990) offer great promise in elucidating the cooperative and group nature of music (Brown in press). Again, music making is not only about within-group cooperation, coordination, and cohesion, but it is principally about these things. How this may relate to the vocalization capacities, group structures, and social behaviors of our hominid ancestors is a matter of central importance for future research and theory in evolutionary musicology.

Acknowledgments

I thank several colleagues who have been generous with their advice and time: Gerhard Schlosser (University of Bremen) for reading the paper and for many stimulating discussions about the nature of homology; Tomas Riad (Stockholm University) not only for his insightful comments on an earlier draft of the paper but for generously spending many hours with me clarifying misconceptions about autosegmental theory and the nature of tone in spoken language; Ulrik Volgsten (Stockholm University) for his critical reading of the paper, and for invaluable discussions about musical semiotics; and Stephen Matthews (Hong Kong University) for many illuminating discussions through e-mail about tone languages. I dedicate this chapter to Mari Mar, Gerhard, and Cristian.

Notes

1. The dichotomy between the acoustic mode and the vehicle mode of music cognition has an important implication for the question of animal song discussed in chapter 1. As I see it, birdsong is not a form of music for exactly the same reason that linguists argue that it is not a form of language. What I call the vehicle mode consists of the representational, iconic, speech-related, and cultural aspects of music, and depends on the rich representational abilities of human beings (see Bickerton 1995). In contrast, when talking about animal song as an acoustic system (analogous to the acoustic mode of human music), it is simply impossible to create a line of demarcation between it and the family of human musics. The vehicle mode is this line of demarcation between music and all forms of nonhuman song.

2. The vehicle mode involves at least seven important functions of music: universal involvement of music in representational rituals; verbal song: songs with words or words with music; music as symbolizer: the use of musical works (or pitches, motifs, melodies, or rhythms therein) to represent cultural objects; music as symbol: extramusical associations of elements of the musical system; acoustic depiction of nonmusical sounds, such as animals, people, and environmental sounds; musical narration: music's use to color actions, events, and characters in the theatrical art forms, such as drama and film; and context switching: reuse of music from one context in another context, for example, classical music in television commercials.

3. The sound emotion system of music consists of at least four major processes: pitch-set effects: contrastive use of different pitch sets (i.e., scales or modes) to convey different emotional meanings; contour-meaning associations: contrastive use of different types of ascending and descending melodic patterns to convey different emotive meanings; blending effects: the emotive effect of sound blends, such as the blendings of pitches (homophony), melodic lines (polyphony), and rhythms (polyrhythms); and progression factors: phrase-level devices for building up coherent and organized musical phrases. In a hierarchical organization of these four components, progression factors sit at the highest level. They are fed into by contour-meaning associations (e.g., ascending and descending melodic lines) and blending effects (e.g., tonicization, cadential formulas, and coordinated motivic movements), which themselves are fed into by pitch-set effects, which contribute factors related to pitch contours, melodic contours, chords, polyphony, etc.

4. One stabilizing selection force that could have kept language tonal during the earlier stages of language evolution was the biological cost in creating anatomical changes to the vocal tract for permitting expansion of the segmental inventory. Evolution of human-specific features of the vocal tract is seen as being essential to the formation of consonants and thus consonant-vowel segments. The capacity to form consonants requires many complex changes in the articulatory mechanisms of the vocal tract, whereas production of several of the vowels can be accomplished even by chimpanzees (de Waal, 1988). Therefore, "it is not a great problem to suggest routes by which at least three distinctive vowels might find their way into the vocal activities of our [hominid] ancestors" (Beaken, 1996:111). The point is that whereas the evolution of new articulatory mechanisms, leading to new consonants, is a costly biological innovation, exploiting pitch contour with vowels is a relatively cheap and simple way of expanding the lexicon. This could have been a major stabilizing selection pressure keeping human language tonal during the earliest stages. One outcome of this reasoning is that intonation languages should have developed, in general, larger segmental inventories than tone languages, as expansion of the segmental inventory

is seen as the key step in reducing the necessity of lexical tone in spoken language. I am indebted to Dr. Stephen Matthews for pointing out to me this putative trade-off between lexical tone and segmental inventory size within languages. As this hypothesis demands the existence of lesser rather than greater sophistication of the vocal tract for speech to occur (fewer rather than more segments), it tends to support theories that call for the early emergence of speech in hominids (see Frayer and Nicolay, this volume).

References

Aiello, R. (1994). Music and language: Parallels and contrasts. In R. Aiello and J. Sloboda (Eds.) *Music Perceptions* (pp. 40–63). Oxford: Oxford University Press.

Beaken, M. (1996). *The Making of Language*. Edinburgh: Edinburgh University Press.

Bernstein, L. (1976). *The Unanswered Question: Six Talks at Harvard*. Cambridge: Harvard University Press.

Bickerton, D. (1995). *Language and Human Behavior*. Seattle: University of Washington Press.

Borchgrevink, H. M. (1991). Prosody, musical rhythm, tone pitch, and response initiation during amytal hemisphere anaesthesia. In J. Sundberg, L. Nord, and R. Carlson (Eds.) *Music, Language, Speech and Brain* (pp. 327–343). Houndmills, UK: Macmillan Press.

Boyd, R. and Richerson, P. J. (1990). Group selection among alternative evolutionarily stable strategies. *Journal of Theoretical Biology* 145:331–342.

Brown, S. (in press). Evolutionary models of music: From sexual selection to group selection. In F. Tonneau and N. S. Thompson (Eds.) *Perspectives in Ethology*, XIII. New York: Plenum.

Clarke, E. F. (1989). Issues in language and music. *Contemporary Music Review* 4:9–22.

Clynes, M. (1977). *Sentics: The Touch of Emotion*. London: Souvenir Press.

Cruttenden, A. (1997). *Intonation*, 2nd ed. Cambridge, UK: Cambridge University Press.

Feld S. and Fox, A. A. (1994). Music and language. *Annual Review of Anthropology* 23:25–53.

Fromkin, V. (Ed.) (1978). *Tone: A Linguistic Survey*. New York: Academic Press.

Goldsmith, J. A. (1976). *Autosegmental Phonology*. Bloomington: Indiana University Linguistics Club.

Goldsmith, J. A. (1990). *Autosegmental & Metrical Phonology*. Oxford: Blackwell.

Goldsmith, J. A. (Ed.) (1995). *The Handbook of Phonological Theory*. Cambridge: Blackwell.

Hood, M. (1971). *The Ethnomusicologist*. New York: McGraw-Hill.

House, D. (1990). *Tonal Perception in Speech*. Lund, Sweden: Lund University Press.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge, UK: Cambridge University Press.

Ladd, D. R. and Terken, J. (1995). Modeling intra- and inter-speaker pitch range variation. *International Congress of Phonetic Sciences 13* (Stockholm) 2:386–389.

Lerdahl, F. (1988). Cognitive constraints on compositional analysis. In J. Sloboda (Ed.) *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition* (pp. 231–259). Oxford: Oxford University Press.

Lerdahl, F. and Jackendoff, R. (1983). *A Generative Theory of Tonal Music.* Cambridge: MIT Press.

Lewin, R. (1993). *Human Evolution: An Illustrated Introduction*, 3rd ed. Boston: Blackwell.

Liberman, M. and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oerhle (Eds.) *Language Sound Structure* (pp. 157–233). Cambridge: MIT Press.

Lomax, A. (1968). *Folk Song Style and Culture*. New Brunswick, NJ: Transaction Books.

Marler, P., Evans, C. S., and Hauser, M. D. (1992). Animal signals: Motivational, referential, or both? In H. Papoušek, U. Jurgens, and M. Papoušek (Eds.) *Verbal and Vocal Communication: Comparative and Developmental Approaches* (pp. 66–86). Cambridge, UK: Cambridge University Press.

Merriam, A. P. (1964). *The Anthropology of Music*. Evanston, IL: Northwestern University Press.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist* 111:855–869.

Morton E. S. (1994). Sound symbolism and its role in non-human vertebrate communication. In L. Hinton, J. Nichols, and J. J. Ohala (Eds.) *Sound Symbolism* (pp. 348–365). Cambridge, UK: Cambridge University Press.

Pierrehumbert, J. (1980/1987). The phonology and phonetics of English intonation. Doctoral thesis, Massachusetts Institute of Technology. Published by Indiana University Linguistics Club (1987).

Pierrehumbert, J. (1991). Music and the phonological principle. Remarks from the phonetician's bench. In J. Sundberg, L. Nord, and R. Carlson (Eds.) *Music, Language, Speech and Brain* (pp. 132–145). Houndmills, UK: Macmillan.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack (Eds.) *Intentions in Communication* (pp. 271–311). Cambridge: MIT Press.

Seyfarth, R. M., Cheney, D. L., and Marler, P. (1980a). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science* 210:801–803.

Seyfarth, R. M., Cheney, D. L., and Marler, P. (1980b). The assessment by vervet monkeys of their own and other species' alarm calls. *Animal Behavior* 40:754–764.

Sloboda, J. A. (1985). *The Musical Mind: The Cognitive Psychology of Music*. Oxford: Clarendon Press.

Sober, E. and Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge: Harvard University Press.

Struhsaker, T. T. (1967). Auditory communication among vervet monkeys (*Cercopithecus aethiops*). In S. A. Altmann (Ed.) *Social Communication Among Primates* (pp. 281–324). Cambridge, UK: Cambridge University Press.

Swain, J. P. (1995). The concept of musical syntax. *Musical Quarterly* 79:281–308.

Swain, J. P. (1996). The range of musical semantics. *Journal of Aesthetics and Art Criticism* 54:135–152.

Swain, J. P. (1997). *Musical Languages*. New York: Norton.

Thomas, D. A. (1995). *Music and the Origins of Language: Theories from the French Enlightenment*. Cambridge, UK: Cambridge University Press.

Thorsen, N. (1980). Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish. *Annual Report of the Institute of Phonetics, University of Copenhagen* 14:1–29.

Thorsen, N. (1981). Intonation contours and stress group patterns in declarative sentences of varying length: Supplementary data. *Annual Report of the Institute of Phonetics, University of Copenhagen* 15:13–47.

Umiker, D. J. (1974). Speech surrogates: Drum and whistle systems. In T. A. Sebeok (Ed.) *Current Trends in Lingustics* Vol. 12 (pp. 497–536). The Hague: Mouton

de Waal, F. (1988). The communicative repertoire of captive bonobos (*Pan paniscus*) compared to that of chimpanzees. *Behavior* 106:183–251.

Wind, J., Chiarelli, B., Bichakjian, B., and Nocentini, A. (Eds.) (1992). *Language Origin: A Multidisciplinary Approach*. Dordrecht: Kluwer.