# 12

# Multiple Levels of Structure in Language and Music

Sharon Thompson-Schill, Peter Hagoort, Peter Ford Dominey, Henkjan Honing, Stefan Koelsch, D. Robert Ladd, Fred Lerdahl, Stephen C. Levinson, and Mark Steedman

**Abstract**

A forum devoted to the relationship between music and language begins with an implicit assumption: There is at least one common principle that is central to all human musical systems and all languages, but that is not characteristic of (most) other domains. Why else should these two categories be paired together for analysis? We propose that one candidate for a common principle is their structure. In this chapter, we explore the nature of that structure—and its consequences for psychological and neurological processing mechanisms—within and across these two domains.

## The Syntax of Music and Language

### A Cautionary Prelude

A theme which runs throughout this book is the importance of recognizing the diversity of forms that are called "music" or "language," and the dangers of an overly narrow focus. Unfortunately, at this stage in the development of these fields, there are limitations in the data that are available for analysis. Therefore, although we have tried to focus on general principles, much of what we have to say about the structure of language is based on written English language and much of what we have to say about the structure of music is based on scores of Western tonal music. The consequences of this limitation for our understanding of structure can be illustrated with one example from linguistics: In contrast to English, Czech is usually described as a language with "free word order." However, word order in Czech acts as a marker for discourse "information structure," marking topic and comment, given and new—a function which intonation performs in English. English employs free order with respect to

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

information structure and is rigid with respect to argument structure, whereas Czech is the reverse. If modern syntactic theory had started with Czech then we might have called English free word order and Czech rigid. In other words, the mere inclusion of cross-cultural comparisons does not ensure a non-ethnocentric approach to the study of language and music structure, in the same way that the study of spoken language comprehension has been completely shaped by the influence of our alphabet-focused written language system.

## Hierarchical Structure

Language and music arrive through the ear and exit through the mouth (or fingers or feet) as actions in time; that is, as a continuous stream of (primarily) acoustic or motor information. But that is *not* the end of the story: we can process an acoustic input by grouping one sound with the next, in the same linear order in which they arrive, with no need to restructure the input. But the situation is quite different. A given linguistic or musical string is best described by a hierarchy, assembled out of elements in the string, in a way that captures meaning relations among the elements beyond their temporal order. Many of the details about the hierarchical organization of elements in music and language (i.e., of syntax) are reviewed by Lerdahl (this volume).

It is important to understand that when linguists talk about hierarchical structure, they distinguish two levels of structure. The most important level of hierarchical structure is the level of *meaning representation*. Such representations are sometimes called "logical forms," because the way linguists write them down often looks like some version of first-order logic, with which it shares such properties as recursivity and compositionality. (This is not to claim that the psychologically real meaning representations look anything like a standard logic.) Such representations are closely related to underlying conceptual relations, standing in a subsumption relation to them, according to which, and at some level, they must be the same for all languages. (The reason for believing this is that it is hard to see how children could learn the language of their culture without access to meaning representations. Since languages differ in their surface forms, and children can learn any of them, this meaning representation must be the same for all.) Since languages differ, in particular, in the order of elements like verbs and noun phrases, we should probably think of logical forms as unordered structures, although of course when we write them down, we will have to choose an ordering on the page. (The fact that so many distinct notations are on offer for essentially the same purpose strongly suggests that the linguists do not have a very clear idea of what the universal logical language really looks like.)

The second kind of structure which linguists talk about is sometimes referred to as *surface structure*. Such structure is a grouping of the elements of the sentence into structural constituents like the English noun phrase and verb phrase. In English and some other relatively rigid word-order languages, such

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

constituents are closely related to such elements of the logical form as predicates and arguments, and can even reasonably be claimed to exhibit a similarly recursive structure. However, other free word-order languages, like Turkish or Latin, do not exhibit any obvious surface constituency and allow considerable freedom for elements that are related at the level of logical form to be nonadjacent in the string. While there is some evidence from phenomena like coordination of some kind of structure, such structure seems to be related to the process of derivation of logical form, rather than to interpretable syntactic structure. In this connection, it is interesting to note that it is commonplace in computational linguistics to regard all surface structure, including that attributed to English, as being epiphenomenal on the process of deriving logical form.

Both in language and in music, elements that are nonadjacent in the sentence may be grouped in the hierarchical meaning representation, as in the case of the "right node raising" construction in (a) language and (b) "interrupted cadences" in music:

(a) I grow, and you sell, beans. = I grow beans, and you sell beans.
(b) $II^7\ V^7$, $II^7\ V^7$, I = $II^7\ V^7$ I, $II^7\ V^7$ I.

One point about these groupings is worth making explicit here: The fact that *grow* belongs with *beans* (and $V^7$ with I) derives from their interpretation. The interest of long-range dependencies is that they show a similarity in the way in which sentences and music are mapped onto meaning; more specifically, sentence-internal semantics and intramusical meaning (see Koelsch, this volume). Whether or not there are other types of meaning is reviewed by Seifert et al. (this volume). Here, we simply make the point that there is a broad similarity between language and music in the way syntax maps strings onto hierarchical meaning representations.

The fact that language and music have structures with nonadjacent, long-distance dependencies reflects a fundamental property of the domain, not some peculiar quirk of each system: We can consider that the semantic content of language is a form of high-dimensional representation. Language production must find a way to transform this high-dimension representation into a linear or near-linear sequence. In making this transformation, items that were "adjacent" (i.e., related) in the high-dimensional space will become separated, thus creating long-distance dependencies. In other words, dimensional reduction (in this case from many to just one) requires a distortion in some of the relations (e.g., compare a map to a globe). Establishing dependencies, including the long-range variety, is the job of syntax.

There may be differences in the types of hierarchical structures (e.g., how deep vs. how flat they are) between music and language, as well as within music and within language. The organization of the levels in these hierarchies is highly culture-dependent: meaning might be expressed in tree structures in a rigid word-order language, like English, and in non-ordered dependency relations in another. Music always seems to use linear order as the fundamental

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

organizing principle, but there are nonetheless examples of nonadjacent dependencies in music.

A related problem that immediately arises in any structural analysis is how to define a constituent; that is, how to extract a discrete object from a continuous signal (i.e., the leaves of the tree). In language, the smallest object (or event) is usually taken to be the phoneme. In music, one can define an event as a separately sounding drum beat, pitch, or chord. It makes sense to simplify the surface of a musical texture to arrive at syntactically useful events without unnecessary clutter, but defining "syntactically useful" is not without its challenges. Consider the "Ooh-ooh-ooh" in the 1935 recording by Billie Holiday that is followed by the beguilingly rhyming "what a little moonlight can do-oo-oo." If you look at the transcription of this opening text, you will see little more than three "ooh's." However, it seems virtually impossible to reduce them meaningfully to individual sounding notes. Where does one note begin and the other end? By contrast, the meaning of "ooh-ooh-ooh," as sung by Billie Holiday, seems to be carried by the (indivisible) whole rather than in the individual sounds and notes. This example demonstrates how letters, and language in general, fall short. Jackendoff and a number of others dismiss melodic utterances like "oh," "wow," "hmm," and "hey" as relics from the "one-word stage" of language. Others emphasize, more plausibly, that such expressions represent a fundamental and very old aspect of language and music.

A second problem that arises concerns the determination of the maximal domain of the hierarchy: For example, in syntax, it is usually presumed to be the clause, but in Western orchestral music it may extend (with perhaps loss of perceptual relevance) to the movement. At the level of maximal domains, one may have a shift of "currency" as it were: a sentence in conversation may constitute part or the whole of a turn; a turn delivers a speech act, which is something more than a chunk of propositional meaning; it counts as an action, so that it can be responded to by an action, verbal or otherwise (cf. requesting the wine, responded to by providing it).

A further set of problems that one might address, but which in our opinion is fruitless, is the effort to relate specific structural elements of music to those of language. There is little to be gained by endless discussion of whether words correspond to notes or something else. Instead efforts should be directed toward clearly formulating a testable hypotheses (e.g., about brain activity; see Koelsch and Patel, both this volume) based on the assumption that there is hierarchical grouping structure in both music and language.

## The Ambiguity Problem

The average length of a sentence in the *Wall Street Journal* is 25 words. If one attempts to compute the meaning of such a sentence from a parser that is drawing from an annotated database of over a million words, one discovers

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

that there are hundreds (if not thousands) of syntactically valid analyses from which to choose. Yet, we can read the *Wall Street Journal* without any difficulty (at least with regard to the parsing problem). The question, therefore, is how humans do this, not only with language but also with music. Here we should distinguish between global ambiguity, as in the sentence *Visiting relatives can be a nuisance*, and local ambiguity, as in a sentence which begins *Have the officers…* (which is locally ambiguous at the beginning of the question or a command) but resolves it at the end, e.g., with *arrived?* or *dismissed!*.

**Ambiguity in Music**

The mapping between the linearly ordered event sequence and the hierarchically organized structure of music is not one-to-one. There is both local and global ambiguity. A diminished chord that contains the note C, as played on the piano, is locally ambiguous in terms of notation: you can write the minor third above the C as an E flat or as a D sharp. If the next chord is G, then it is a "C diminished" chord and you express the tone as an E Flat. If, however, the piece is in E minor, then it is an "A diminished" chord and you write the tone as a D sharp. In the "whole-tone" scale used by Debussy, music is based on the augmented chord, which is similarly ambiguous to the diminished, but in whole-tone music, it never gets disambiguated. As a result, many of Debussy's pieces are in a sense globally ambiguous as to any tonal center or key. Thus, just as the reader of the *Wall Street Journal* needs to extract one interpretation from many possibilities, so too does the listener to music.

In light of the pervasive ambiguity in music, several principles describe how one interpretation comes to be favored over another. These principles describe transitions in a multidimensional tonal space that is crystalline in its multidimensional regularity and beauty. We briefly digress from the topic of ambiguity to describe this space (another "structure" of music).

The development of pitch space models began in pedagogical eighteenth-century music theory treatises. Part of the cognitively valid solution was intuitively achieved already in the early eighteenth century (Euler 1739), and was developed in computational terms by Longuet-Higgins and Steedman (1971), using a Manhattan city-block distance metric for harmonic distance. In the early 1980s, the experimental psychologist Krumhansl and collaborators established empirically the shape of tonal spaces at three levels of organization: pitch, chord, and key (Krumhansl 1990). In Lerdahl's *tonal pitch space*, he develops a quantitative music theoretic model that correlates with the data and unifies the three levels formally (Lerdahl 2001b). Lerdahl and Krumhansl (2007) successfully tested the tension model that relies in part on the pitch space theory.

When selecting an interpretation of an ambiguous musical event, the principle of the shortest path ranks and selects the most efficient, most probable solution to both tonic-finding (finding the tonal point of reference) and

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

tree-building (forming a hierarchical representation of events). It does this by measuring distances in the tonal space. The attraction component treats the variable tendencies of pitches to move to other pitches. For example, the leading tone is strongly attracted to the tonic, which is more stable and very proximate. The relation is asymmetric: the tonic is only weakly attracted to the leading tone. Attractions are calculated for each voice in a harmonic progression, and the shortest path also enters into the equation. The overall picture is a kind of force field within which pitches and chords behave: the stronger the attraction, the stronger the expectation. As with the principle of shortest path, this procedure can be cast in terms of probabilities. All but the strongest probabilities are pruned away quickly. If an improbable event follows, the experience is a surprise or jolt. The principle of prolongational good form supplements the principle of the shortest path in building a tree structure for event sequences. Prolongational good form encourages, among other things, the characteristic tonic–dominant–tonic (I–V–I) relationship that is at the heart of classical tonal music. Typically, this relationship occurs recursively in the course of a piece.

Thus far we have talked only about pitches and chords, but as Lerdahl reviews (this volume), there are also hierarchical rhythmic structures (although in the case of rhythm, groups do not form dominating-subordinating constituencies). The mapping of rhythm events onto a hierarchy is also ambiguous. As an example, consider a rhythm of three time intervals: 0.26 s to 0.42 s to 0.32 s (i.e., which occur on a continuous timescale). If that rhythm is primed (preceded) by a rhythm in duple meter, it will be perceived by the majority of (Western) listeners as 1:2:1. However, when the same rhythm is preceded by a fragment of music in triple meter, then the majority of participants will perceive it as 1:3:2. Physically, the musical events are identical. However, perceptually and cognitively they are distinct; this turns out to be common in the space of all possible rhythms of a certain duration (Desain and Honing 2003). Research in rhythmic categorization has shown that this process remains open to top-down cognitive influences, either influenced by the preceding musical context (veridical expectation) or by expectations constructed from earlier exposure to music (schematic expectation) (Bharucha 1994; Huron 2006). A consequence of this is that hierarchical analysis based on categorized rhythm (e.g., $16^{th}$–$8^{th}$–$16^{th}$ notes or 1:2:1; cf. Lerdahl, this volume) is dependent on the outcome of the analysis of which it is actually the input.

## Ambiguity in Language

As mentioned above, a computational model of parsing based on corpus data shows a remarkable degree of syntactic ambiguity. As Steedman has noted, "the reason human language processing can tolerate this astonishing degree of ambiguity is that almost all of those syntactic analyses are semantically completely anomalous." Thus, the resolution has to come from the interfaces with discourse semantics and world knowledge, but how these interface operations

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

are computationally handled in an incremental processing system is unsolved. (We return to a longer discussion of this unsolved problem at the end of this chapter.)

It is interesting to consider the use in language of "least-effort" heuristics, of the kind that were applied to musical disambiguation in the last section. Similar "shortest move" principles have frequently been proposed in linguistics since Rosenbaum's Minimal Link Condition on control (Rosenbaum 1967), most recently in the form of the economy principles of Chomsky (1995). Such principles were proposed in answer to the question, "Why does the long-distance dependency upon the subject of the infinitival *to go* in the sentence *Mary wants John to go* refer to John's departure, not Mary's, as it would be in the sentence *Mary wants to go*? They claim that it is because, in both cases, the infinitival must choose the closest antecedent noun phrase. The trouble is that there is a small class of verbs, like promise, whose infinitivals target the nonproximal noun phrase, as in *Mary promised John to go*, in which it is Mary's departure that is at stake.

Earlier we said the sole *raison d'etre* of syntax is to build structural meaning representations, and that surface structure should be viewed as a record of the process by which the meanings get built. It follows that the operations of surface syntax give us something on which to hang the Bayesian priors of a parsing model. Such parsing models have to disambiguate quickly, as we do not have the luxury of contemplating thousands of possible structures before we select the most likely one. As Levinson's turn-taking work illustrates (this volume), we have to have disambiguated an utterance before it is finished (in order to plan our own).

However, if the point of a parsing model is to disambiguate, why are both music and language not merely ambiguous systems, but are designed to yield massive numbers of irrelevant parses? Some have argued that grammar-based parsing models really play a very limited role in comprehension and exist primarily to regularize production (Ferreira 2007). That is, syntax might exist for production but be relatively useless for comprehension. Of course, this creates a new puzzle; namely, how do we get to semantics without syntax and what analysis is "good enough" for comprehension, without requiring a full or correct parse?

Instead of questioning whether or not comprehension requires parsing, another approach to this puzzle is to question some of the assumptions that go into the models. The syntax–semantics interface seems to work quite well in human language comprehension, but appears to raise severe problems for machine processing of language. This is partly because extrasyntactic sources of information (e.g., context, world knowledge) are known to play an important part in disambiguation. Still, it is currently difficult to exploit such knowledge in computational systems, due to the lack of adequate semantic representations.

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

This suggests that the current separation and representation of syntax and semantics may have some fundamental problems.

## Ambiguity between Music and Language

An initial question was posed at this Forum: If an archeologist from another era were to come across several music scores, how could this be determined to be musical notation and not fragments of a written language? Alternatively, how could samples of written language be classified as text and not musical notation?

We asked a slightly different question: When a listener hears an acoustic signal (with a certain set of spectral properties), how does the listener decide whether it is language or music? This represents an additional type of ambiguity, which occurs not at the level of mapping events onto objects in a hierarchical structure, but at the level of constructing the events in the first place.

One might object, at this point, and argue that this is not a type of ambiguity that occurs unless one is trying to discriminate between forms of language and music that are closer to the center of the music–language continuum. However, to illustrate that even an English speaker familiar with Western tonal music can have two incompatible interpretations of a single acoustic input, we describe an illusion first observed by Diana Deutsch (1995): A sentence containing the phrase "sometimes behave so strangely" is perceived by a listener as intended (i.e., as a sentence). If this snippet of the phrase is looped repeatedly, perception changes. As semantic satiation takes hold, the phrase begins to sound like music. Indeed, when the phrase is heard again, replaced in the middle of the sentence from which it was removed, it continues to be perceived as song in the middle of an otherwise normal sentence.

This illusion illustrates a point that is both obvious and profound. The obvious part is that both language and music share a sensory channel, which begins at the cochlea. As such, this allows for the possibility of competition between interpreting the signal as speech versus music. Just as a Necker cube cannot be interpreted as being in two orientations at once, so too must the listener select a single interpretation of "sometimes behave so strangely" (and other such phrases that produce this illusion). Moreover, if the interpretation is as language, high pitch is heard as stress or accent, but if the interpretation is as music, high pitch is heard as unaccented pitch (Ladd, pers. comm.). The profound part is that the illusion reveals that there is not sufficient information in the signal itself to discriminate unambiguously between these two interpretations. The "decision" appears to be made on the basis of something that is not acoustic at all; namely, semantics. This is not to say that, on average, the acoustic signal between music and language does not differ (which of course it does) but rather that there is overlap between their distributions. Information

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

is extracted from the acoustic signal that allows one to select the most likely interpretation. Next we consider the streams of information, extracted from an acoustic input, that are used to construct the objects of analysis in linguistic and musical structures.

## Streams of Information

Our discussion of structure has thus focused on one particular structural description: the hierarchies that are constructed out of a linearly ordered sequence of events, such as phonemes/words or notes/chords. Here we turn to a different type of structural analysis: a *description of the system* (i.e., the types of information in the signal that are represented) as opposed to a description of the content itself (i.e., the structure of a phrase). This is a nonhierarchical structure that we will henceforth refer to as a *set of streams*, rather than by hierarchically ordered levels (as with syntax). These streams can be partially independent (unlike syntactic trees) and can be used for different functions. In the simple case, a stream of information can be thought of as a distinctive modality or medium of transmission, like manual gesture which accompanies speech. However, even within a modality, there are different types of information that must be extracted from a single acoustic signal.

### The Big Three

When linguists refer to types of information, or representations, in language, they are often referring to *semantics*, *syntax*, and *phonology*. We, too, could have approached the question of the structure of the language system with these terms but chose not to for several reasons. First, the distinctions between these three domains in language are not entirely clear. Although they are necessary constructs for linguistic theory, it is not clear that they are distinct kinds of representations or processes; this may explain why efforts to localize "syntax" or "semantics" to a discrete cortical module have by and large been unsuccessful. Second, in the context of this Forum (the relation between language and music), analyzing the semantics, syntax, and phonology of language immediately invites comparisons of each of these subsystems to some counterpart subsystem in music. Just as trying to relate notes to words is fruitless, so too is the attempt to match parts of the language system to parts of the music system. Semantics, syntax, and phonology are functional descriptions, and because the functions of language and music are different, it is hopeless to impose one system's labels on the other system. As we reviewed above, there is, however, overlap in the processing of language

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

and music at levels of analysis where we can make comparisons; namely, the content of the signal. One can usefully ask what kind of information can be extracted from the acoustic signal, what functions each type of information supports (within language and music), and what similarities and differences exist between the types of information used—and the means by which they are integrated—in these two domains.

## The Problem of Prosody

One example that illustrates the difference between using a functional label and a description of the information that a function requires is in the domain of prosody. Ladd (this volume) discusses the confusion that has been created by the catchall use of the word prosody to mean "everything left in language when you remove the words." This conventional (albeit recent) use of the word prosody as a functional label has the effect of implying the same function for a whole host of different "suprasegmental" signals. As reviewed in his chapter, the clearest case where this coarse functional grouping is inappropriate is that of lexical tone, which plays a purely phonological role; that is, tone variation in Mandarin has the same function as voice onset time variation in English. (Rather than attributing the error of calling lexical tone prosody to sloppiness, we suspect this is another example of the English-centric influence on linguistics.)

We believe that we can clear up these muddy waters by shifting the emphasis from theoretically laden functional labels to more neutral informational descriptors. There is something in common to "everything but the words"; namely, that unlike the words (i.e., unlike consonants and vowels which require discrimination of very rapid transitions in the acoustic signal), lower-frequency information is used to discriminate lexical tone (in a tonal language), accent (in a nontonal language), intonation and phrasing (at the sentence level), and emotional content. In turn, these discriminations can affect phonology, syntax, and semantics.

We also believe that descriptions of the content of the acoustic signal aids in the interpretation of observed similarities and differences between language and speech. Below, we discuss some of these comparisons (e.g., lateralization differences).

## Decomposing the Signal

If we are to make any progress in understanding the different kinds of information which are present in an acoustic signal that support the functions necessary for language and music processing, we must solve the problem of how a single signal is decomposed into its constituent parts. If the parts are, as suggested in the discussion of prosody, distinguished from each other based

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

on their temporal frequency in the signal, all that is needed is a filter with a specific bandwidth. Hagoort and Poeppel (this volume) describe a candidate mechanism for implementing such a system. This mechanism for chunking speech (and other sounds) is based on a neuronal infrastructure that permits temporal processing in general. In particular, intrinsic cortical oscillations at different frequency bands (e.g., theta between 4–7 Hz, gamma > 40 Hz) could be efficient instruments for decomposing the input signal at multiple timescales. Neuronal oscillations reflect synchronous activity of neuronal assemblies. Importantly, cortical oscillations can shape and modulate neuronal spiking by imposing phases of high and low neuronal excitability (cf. Schroeder et al. 2008). The assumption is that oscillations cause spiking to be temporally clustered. Oscillations at different frequency bands are then suggested to sample the speech signal at different temporal intervals. There is evidence that left temporal cortex auditory areas contain more high-frequency oscillations (closely corresponding to the length of the rapid transitions of consonants and vowels) and that right temporal cortex auditory areas are more strongly dominated by more low-frequency oscillations (closely corresponding to the length of syllables). In addition, we know that auditory signals with high-frequency patterns produce more activation in the left temporal cortex, whereas low-frequency patterns produce more activation in the right temporal cortex.

According to this account, information from the acoustic signal (for speech or music) is decomposed into (simplifying here, to only two streams) high- and low-frequency information. If the acoustic stream is speech, high-frequency information can be used to discriminate phonemes, whereas low-frequency information can be used to calculate stress, accent, or (in a tonal language) tone. If the acoustic stream is song or music, there is less information present at high frequencies (music, including song, is slower than speech on average), which might explain the relative prominence of right over left temporal activation during music compared to speech perception. In effect, Poeppel's suggestion is that that the biological constraints (i.e., the oscillations that are part of the "hardware") on speech comprehension may have shaped the properties of our language, capitalizing on these naturally occurring oscillation frequencies to split the signal into what we now call phonemes and syllables.

Music, of course, would be analyzed with the same streams, segregated by the same oscillatory mechanisms. However, in music, quite different timescales operate: low-frequency scales can be associated with the pulse (or tactus) of the music (in the order of 400–600 ms). A cognitive phenomenon named beat induction is commonly associated with brain regions such as the basal ganglia (Grahn and Brett 2007). There are also faster timescales associated with variable durations (i.e., rhythm, associated with activity in the cerebellum; Grube et al. 2010) and expressive timing (minute intentional variations in the order of 50–100 ms). Finally, on a comparable timescale, there are timbral aspects of music, such as the information that human use to distinguish between instruments from the attack of the acoustic signal.

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

If spreading information across the channels is useful for language comprehension, why then does music not capitalize on this split? One provocative idea is that it used to, under theories that music originated from song (so the high-frequency channel would have used for phoneme discrimination too). Physical constraints may also factor into how long it takes to produce a specific pitch (when, as in music, a rough approximation is not acceptable).

Although this whole argument sounds very tidy, if not simplistic, it has problems. First, the role of intonational marking of phrase boundaries is not immediately clear. A lot of the evidence from recent work on intonation indicates that, to a very considerable extent, intonation works in terms of local pitch events, not holistic contours. How those get produced and interpreted does not seem to fall out from thinking in terms of smaller and larger domains. (Perhaps an intonational boundary marker is similar to a chord change: the new chord itself is a local event, but it defines a new larger stretch of the harmonic structure.)

One potential consequence of decomposing a signal into separate streams so early in processing is that there has to be some mechanism for maintaining coordination between the streams in tight temporal alignment. It is not clear how best to think about this. It seems inappropriate to treat the separate streams as separate in the sense of Bregman (1990): it is well established that syllable-level pitch movements are very precisely aligned relative to the articulatory gestures for consonants and vowels (e.g., Arvaniti et al. 1998), and that (unlike in Bregman's research) listeners are very sensitive to differences in alignment. Further research is required to reconcile the apparent separateness of the streams from the equally apparent unity of the whole signal.

## Cross-Modal Streams

Our discussion of streams of information has thus far focused on decomposing the acoustic signal. However, the coordination problem just raised extends to other modalities as well. We will use as our case study of cross-modal integration the case of gesture. Of course, there is one population in which gesture and language occupy a single modality; namely, users of sign language. Interestingly, some of the issues raised above are pertinent to studies of sign language as well, such as the location of information differing in functional relevance on the face.[1]

There are a number of different categories of gestures, including some that are tightly locked to the onset of a word (e.g., indexical gestures, such as pointing to accompany "this one"). Others precede the onset of a word by only a fractional period of time (e.g., an iconic gesture, such as a hammering

[1] When thinking about the relation between emotion and language and speech, it may be worth considering whether emotion is more like a modality (e.g., hand movements) or a functional system (e.g., like prosody).

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

motion when saying "hammer"). Despite these slight timing differences, both the speech signal and the co-speech gestures result in a common representation, and hence have to be integrated in comprehension and jointly planned in production. Frontal cortex seems to play a role in this integration process (Willems et al. 2007). In ordinary conversational settings, even when speakers talk on the phone, speech does not occur in isolation but is embedded in a multimodal communicative arrangement, including the use of hand gestures.

## Where Next?

### Linking Psycho- and Neurolinguistics with Computational Linguistics

Currently, syntactic processing in computational linguistics and psycholinguistics or neurocomputational models of human sentence processing have almost entirely diverged, and pay almost no attention to each other in the literature. The reason is that parsing on the scale that is required to read the newspaper requires very large grammars, with thousands of rules, and that very large grammars engender huge ambiguity, with hundreds and even thousands of valid derivations for each sentence. Accordingly, state-of-the-art parsers use statistical models of derivations, which allow a probability to be assigned to any analysis of a given sentence. Such statistical models are derived from human-labeled sentences in a "treebank."

These models are rightly despised by psychologists and, in fact, are quite weak, working at about 90% accuracy on a number of measures. They are weak because we do not have enough labeled data on which to train them (and we never will). Psychologists know that the parser draws on all levels of linguistic representation for disambiguation, incrementally and at high speed, including semantics, and even referential context and logical inference. One might expect that they would be able to offer an alternative to the computationalists.

Unfortunately, the models that psycholinguists currently embrace seem to be predicated on the assumption that you sometimes have at most two alternatives, and propose strategies such as "best-first" which have no chance at all of coping with realistically large levels of ambiguity. Moreover, all semantic theories on offer from linguists exhibit highly complex mappings to syntactic structure, involving processes like "covert movement," with which it is very hard to do effective inference. Part of the problem, as Levinson's work shows, is that whatever the real semantics is, the markers found in real languages do not seem to be transparent to the primitive concepts of the presumed universal semantics. Indeed, it seems possible that there is no such primitive that is transparently marked in any attested language.

The open problem that our discussions raised is this: Can we provide psychologically plausible parsing mechanisms that will work at the scale of real human language processing, and can we identify a "natural" semantics and

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

conceptual system which supports inference that can be smoothly integrated with them?

## Musical Dialog

Levinson (this volume) stresses that a central aspect of language is the online dialogic interaction: while one is listening to a partner, one is simultaneously predicting the partner's upcoming words and preparing one's own subsequent utterance. This requires three concurrent but distinct linguistic representations to be managed at one time: (a) the representation of what the speaker is saying, (b) the representation of what the listener believes this speaker will say (which we know from Levinson is often the same as the former but which still must be tracked), and (c) the representation of the listener's planned utterance which is being prepared. These representations, of course, are being crafted in the face of all of the ambiguity just discussed. From a processing perspective, this parallel comprehension and production is very distinct from "passive" listening to a narrative. There are conversations in some musical forms, although there may be some differences between the demands they create in music and in language, depending on whether one musician is creating a plan in response to the music of the other musician or not. Just as conversational turn-taking has processing implications in language, the study of musical dialog may constrain hypotheses about musical production and comprehension. It may also be fruitful to examine the extent to which there are common mechanisms in language and music for managing these interactions.

## Physical Constraints

Above we suggested that properties of the events that compose language and music reflect biological constraints, such as the proposed correspondence between the length of a syllable and the theta oscillation. There are other kinds of biological constraints that one might also usefully consider. For example, in language, the need to breathe constrains the length of a prosodic utterance. The musical analog is the "phrase," which has about the same length as a prosodic utterance. A typical phrase ends in a cadence (formulaic way of achieving closure within a given musical style). The analog is close because much, if not most, music is sung. The breathing constraint applies not only to the voice but also to wind instruments (woodwinds or brass), though not to string or percussion instruments. Nevertheless, music played by winds usually follows the same phrase lengths as vocal music.

Of course, a tempting correspondence such as this may prove to be misleading. Although there may be some evolutionary link between breath groups and linguistic phrases, as between opening and closing jaw gestures and syllables, there is a lot you can say about syllable structure for which basic oscillatory

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

jaw movement is essentially irrelevant. The same may be true of phrases and breath groups.

## Action

Cross et al. (this volume) discuss the problem of the evolution of language and music. To their thoughts, we add an insight that comes from our consideration of the structure of language and music. As described above, both language and music require mapping between a linear sequence and a hierarchical structure, which may involve grouping events that are nonadjacent but which are, instead, connected by their meaning. So, too, elementary actions can be sequenced to form compound actions or plans that have a hierarchical structure. Sensorimotor planning of this kind, including planning that involves tools, is not—unlike language—confined to humans. The mastery of the relevant action representation, including tool use and effects on other minds, also immediately precedes the onset of language in children.

Planning in nonlinguistic and prelinguistic animals is striking for two reasons: (a) the ability to sequence actions toward a goal in abstraction from their actual performance and (b) the fact that this ability is strongly dependent on an affordance-like association between the immediate situation and the objects that it includes, and the actions they make possible. The close relation between planning with tools and other minds and language suggests that this kind of planning provides the substrate onto which language can be rather directly attached, both in evolutionary and developmental terms.

The field of artificial intelligence has created computationally practical representations of actions and planning. It might be interesting to consider how linguistic syntax and semantics (as well as aspects of nonsyntactic speech acts related to discourse, discussed by Levinson, this volume) could be derived from such representations.

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.