17

The Neurobiology of Language, Speech, and Music

Jonathan Fritz, David Poeppel, Laurel Trainor,
Gottfried Schlaug, Aniruddh D. Patel,
Isabelle Peretz, Josef P. Rauschecker, John Halle,
Francesca Stregapede, and Lawrence M. Parsons

Abstract

To clarify the domain-specific representations in language and music and the common domain-general operations or computations, it is essential to understand the neural foundations of language and music, including the neurobiological and computational "primitives" that form the basis for both, at perceptual, cognitive, and production levels. This chapter summarizes the current state of knowledge in this burgeoning cross-disciplinary field and explores results from recent studies of input codes, learning and development, brain injury, and plasticity as well as the interactions between perception, action, and prediction. These differing perspectives offer insights into language and music as auditory structures and point to underlying common and distinct mechanisms and future research challenges.

Introduction

We seek to characterize the *relations* between language and music—and their neurobiological foundations—in the hope that this will lead to real advances and a unified understanding of both. For example, such cross-domain research may elicit change in the architecture of music theory and the psychology of music; deepen our insight into the roles of vocal intonation, cadence, rhythm, rhyme, inflection in emotional expression in drama, poetry and song; shed light on the ontogeny of music and language development in children; reveal common cognitive operations in language comprehension and music perception; and clarify the specialization and the relative contributions of the right and left hemispheres to these two domains. With these possibilities in mind, our

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.

discussions summarized in this report centered on two issues: First, how can we be precise and explicit about how to relate and compare the study of language and music in a neurobiological context? Second, what are the neural foundations that form the basis for these two dimensions of human experience, and how does the neural infrastructure constrain ideas about the complex relation between language and music? For example, to what extent is neural circuitry shared between music and language, and to what extent are different circuits involved? More ambitiously, we ask at what level of analysis can we go beyond "correlational"' statements (e.g., "brain area X underpins music function Y") and strive for accounts of the underlying processes.

An important caveat before we proceed: Speech is propositional, unlike music, and speech sounds carry denotative meaning, unlike most music. However, music can evoke and elicit emotion directly without linguistic mediation. Thus, while we explore the similarities, we must also be aware of the differences between the two domains. Moreover, only a very restricted set of issues is addressed here: one that reflects the composition and expertise of our working group as well as the actual content of the discussions. By design, other chapters in this volume address the issue of structure and syntax (Thompson-Schill et al.) and meaning/semantics in music and language (Seifert et al.), with evolutionary considerations (Cross et al.). Here, we focus on the input interfaces (i.e., perception of music and speech), the relation to production or action, and questions about their neurobiological implementation. However, even within this narrow scope, there are productive areas of study that, regrettably, do not receive the attention they deserve, including the study of song and dance as possible approaches to explore the relations between music, speech, language, and the brain. We leave these important topics for further discussion and research (see, however, Janata and Parsons, this volume).

By way of overview, the chapter proceeds as follows:

1. We outline an approach to characterizing the domains of language and music (from a perception–action perspective), with the goal of identifying tentative lists of basic, fundamental elements, or "primitives," in music and language and discussing how they might be related across domains. We pursue the hypothesis that there are domain-specific representations in language and music and domain-general operations or computations, and we discuss some candidate areas, such as sequencing and attention in time.
2. We discuss some cross-cutting issues that have been investigated productively in both domains and that illustrate areas in which the brain sciences add considerable insight, including the nature of the input codes, learning and development, brain injury and plasticity, and the perception–action cycle. Anticipation-in-time or prediction is an

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.

essential component of the sequencing process in language and music, and will be considered as part of a perception–prediction cycle.

3. Although we point to neurobiological data throughout the chapter, the section on neurobiological constraints and mechanisms focuses explicitly on some of the neural mechanisms that are either reasonably well understood or under consideration. Building on the previous sections, we distinguish between data pointing to domain-specific neural correlates versus domain-general neural correlates.
4. In the final section, we return to some of the open questions, pointing especially to the important contributions that can be made by work on dance, music, poetry, and song. However, the overall, more modest goal of the report is to highlight a set of experimental approaches that can explore the properties and neural bases of the fundamental constituents in music and language.

Structure of the Domains: Computational Primitives and Processes

To identify the relations between domains, it can prove useful to characterize the problem in two ways. One approach is to (attempt to) spell out for each domain an "elementary parts list." For example, such a list for language might include hypothesized representational primitives (e.g., phoneme, syllable, noun phrase, clause) or operational/processing or computational primitives (e.g., concatenation, linearization, dependency formation). Similarly, for music, a list might include tone (representation) and relative pitch detection (computation). The decomposition into constituent parts is necessarily incomplete and the list (Table 17.1) changes as progress is made. For example, such inventories of primitive elements will be modified as one considers the relation of music to dance or speech to song.

A second approach derives from the work of computational neuroscientists such as David Marr, who provided a way to talk about the characterization of complex systems, focusing on vision (Marr 1982). In a related perspective, Arbib (1981) employed schema theory to chart the linkages between perceptual structures and distributed motor control. The neural and cognitive systems that underlie language and music are usefully considered as complex systems that can be described at different levels. At one level, *computational goals and strategies* can be formulated, at an intermediate level the *representations and procedures* are described, while at a third level lie the *implementation* of the representations and procedures. It is assumed that these distinct levels are linked in principle, although actual close linkages may not always be practical. Commitments at one level of description (e.g., the implementation) constrain the architecture at other levels of description (structural or procedural) in principled ways.

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.

Table 17.1 Levels of analysis: A Marr's eye view.

Implementational (domain general)	Hypothesized *implementational* (*neurobiological*) *infrastructure* • Generic forms of circuitry • General learning rules which can adapt circuits to serve one or both domains	
Algorithmic computational (domain general)	Hypothesized *computational primitives* • Constructing spatiotemporal objects (streams, gestures) • Extracting relative pitch • Extracting relative time • Discretization • Sequencing, concatenation, ordering • Grouping, constituency, hierarchy • Establishing relationships: local or long distance • Coordinate transformations • Prediction • Synchronization, entrainment, turn-taking • Concurrent processing over different levels	
Representational computational (domain specific)	Hypothesized *representational primitives*: *language* • Feature (articulatory) • Phoneme • Syllable • Morpheme • Noun phrase, verb phrase, etc. • Clause • Sentence • Discourse, narrative	Hypothesized *representational primitives*: *music* • Note (pitch and timbre) • Pitch interval (dissonance, consonance) • Octave-based pitch scale • Pitch hierarchy (tonality) • Discrete time interval • Beat • Meter • Motif/theme • Melody/satz • Piece

Pursuing the Marr-style analysis, at the top level (i.e., the level of overall goals and strategies) we find perception and active performance of music (including song and dance) as well as language comprehension and production. At the intermediate level of analysis, a set of "primitive" representations and computational processes is specified that form the basis for executing "higher-order" faculties (Table 17.1). At a lower level of analysis is wetware: the brain implementation of the computations that are described at the intermediate level. What may seem a plausible theory at one level of analysis may require drastic retooling to meet constraints provided by neural data. A special challenge is provided by development. Since the brain is highly adaptive, what is plausibly viewed as a primitive at one time in the life of the organism may be subject to change, even at relatively abstract conceptual levels (Carey 2009),

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 the organism may bootstrap, building perceptual and cognitive algorithms, which may form building blocks for subsequent processing in the adult.

Decomposing/Delineating Elementary Primitive Representations and Computations

The richness and complexity of music and language are well described elsewhere in this volume (see, e.g., chapters by Patel, Janata and Parsons, as well as Hagoort and Poeppel). Here, we attempt to identify domain-specific and domain-general properties of music and language, with special (but not exclusive) reference to how constituent processes may be mapped to the human brain and, in some cases, related to mechanisms also available in the brains of other species. Table 17.1 provides the overview of how these questions were discussed, and what kind of analysis and fractionation yielded (some, initial) emerging consensus.

Domain-Specific Representational Inventories: The Primitive Constituents, or "Parts List"

Though there will inevitably be disagreements about the extent to which a given concept constitutes a basic, primitive unit, we offer some candidates for a representational inventory that may underpin each domain—representations without which a successful, explanatory theory of language processing or music processing cannot get off the ground. Regardless of one's theoretical commitments, it would be problematic to develop a theory of language processing that does not contain a notion of, say, syllable or phrase; similarly, a theory of music that does not refer to, say, note or meter will most likely be irreparably incomplete.

If we focus, to begin, on language in terms of the input sound patterns of speech (and forget about written language and orthography altogether), then we seek primitives required for phonetic featural representation. These acoustic and articulatory distinctive features (Stevens 2000; Halle 2002) form the minimal units of description of spoken language from which successively higher levels of linguistic representation are derived. At the most granular level, the speech system traffics in small segment-sized (phonemic) as well as slightly larger syllable-sized units. These need to be "recovered" in the context of perception or production. Segmental and syllabic elements are combined to form morphemes or roots or, more colloquially, words; the combinations of sounds forming words are subject to phonetic and phonological constraints or rules. Although this chapter focuses on spoken language, we find that the lower-level units of signed language are somewhat different and that higher levels (say words) converge.

One of the remarkable features of language is the large number of morphemes (or words) that are stored as the "mental lexicon." Speakers of a language have

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.

tens of thousands of entries stored in long-term memory, each of which they can retrieve within ~200 ms of being uttered (given standard speaking rates). This feat of memory is impressive because the words are constructed from a relatively small set of sound elements, say in the dozens. This could plausibly lead to high confusability in online processing, but the items and their (often subtle) distinctions (e.g., /bad/ vs. /bat/ or /bad/ vs. /pad/) are stored in a format or code that permits rapid and precise retrieval. The words are sound–meaning pairings, often with rather complicated internal morphological structure (e.g., an un-assail-able pre-mise), but aspects of sound and meaning may be distributed across multiple brain regions, with strong associative links between them. Combining these elements pursuant to certain language-specific regularities (syntax) yields phrases and clauses (e.g., "the very hungry caterpillar") that yield compositional meaning. Ultimately, the information is interpreted in some pragmatic or narrative discourse context that provides common ground about knowledge of the world and licenses inferences as well as being integrable into ongoing conversation (Levinson, this volume).

The interrelations among levels are currently being investigated in linguistics and psycholinguistics (see Thompson-Schill et al. and Hagoort and Poeppel, both this volume) and fall outside the scope of this report. However, it is worth bearing in mind that when even a single sentence or phrase is uttered, all of these levels of representation are necessarily and obligatorily activated across multiple brain regions. It is, by contrast, less clear to what extent the same consequences obtain in the musical case for nonmusicians. When speech and language are used in even more multimodal contexts—say during audiovisual speech, spoken poetry or during singing—further levels are recruited, including visual representations of speech and musical representations during singing (and, presumably, motor representations during both types of actions).

Turning to music (Table 17.1, bottom right column), similar categories apply to some degree to the lower levels of speech and musical structure; namely (and minimally) grouping, meter, and pitch space. The assignment of all of these categories involves a complex interaction of primary sensory cues. Thus the constituents of music, which we will refer to, following Lerdahl and Jackendoff (1983), as the musical group, can be induced by alterations in pitch, duration, and, to a lesser degree, amplitude and timbre. Of these, duration is arguably the most determinative, with boundaries tending to be assigned between events that are relatively temporally dispersed. However, if we broaden our study of music from sound patterns to dance, then further primitives are required to link motion of the body with musical space. Moreover, music is often an ensemble activity (e.g., dancing together, singing together, an orchestral performance; see Levinson and Lewis, both this volume) which forces us to assess how the primitives within an individual's behavior are linked with those of others in the group.

Metrical structure—the periodic alternations of strong and weak temporal locations experienced as a sense of "beat"—is also largely induced by

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.

manipulation of these parameters. Length, again, is probably the most determinative, with a somewhat reduced role for pitch. (Indeed, some highly beat-oriented music, such as some forms of drumming, may lack pitch altogether.) Here, amplitude and timbre, insofar as these create a musical accent, tend to be highly determinative. Finally, with respect to harmony–pitch space, pitch is required to be primary, with listeners orienting themselves in relationship to a tonic, perceptual reference point according to which other pitch categories are defined. While Western functional harmony makes use of an extremely rich system of pitch representation, the great majority of musical styles make use of a scale, allowing for a categorical distinction between adjacent and nonadjacent motion with respect to the tonal space (steps and skips). Perhaps most prominently, a sense of closure or completion resulting from the return to a tonic is a recurrent, if not universal, property of Western musical systems.

By analogy to the language case, the music theorists and musicians in our working group guided the discussion and converged on a list of primitives in three parts: spectral and timbral elements related to pitch and pitch relations (note, pitch interval, octave-based pitch scale, pitch hierarchy, harmonic/inharmonic, timbre and texture), elements related to time and temporal relations (discrete time interval, phrase, beat, meter, polyrhythms), and elements related to larger groups (motif/theme, melody/satz, piece, cycle). Here, too, there is a hierarchy of elements. By analogy to language, exposure to a melody presumably entails the obligatory recruitment of the elements lower in the hierarchy, such as the temporal structure of the pitch-bearing elements.

Comparison of Primitives

It is, of course, tempting to draw analogies and seek parallels. Indeed, if one focuses on the lower, input-centered levels of analysis (phoneme, syllable, tone, beat, etc.), one might be seduced into seeing a range of analogies between music processing and the processing of spoken language (including, crucially, suprasegmental prosodic attributes such as stress or intonation). However, if one looks to the representational units that are more distal to the input/output signal (e.g., morphemes, lexical and compositional semantics) in language, possible analogies with music become metaphorical, loose, and sloppy. In fact, closer inspection of the parts lists suggests domain specificity of the representational inventories.

Let us briefly focus on some differences. One general issue pertains to whether (knowledge and processing of) music can be characterized in a manner similar to language. For example, sound categorization skills in language and music may be linked (Anvari et al. 2002; Slevc and Miyake 2006). Although the approach is too simple, a useful shorthand for discussion is that language consists, broadly, of words and rules (Pinker 1999); that is, meaning is created or interpreted by (a) looking up stored items and retrieving their attributes or individual meanings and (b) combining stored words according

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.

to some constraints or rules and generating—or "composing"—new meaning. Both ingredients are necessary, and much current cognitive neuroscience of language has focused on studying how and where words are stored and how and where items are combined syntactically and semantically to create new meaning (for a review, see Hickok and Poeppel 2007; Lau et al. 2008; Hinzen and Poeppel 2011; for a recent meta-analysis of word representation, see DeWitt and Rauschecker 2012).

One question that arises is whether there exists such a construct as a "stored set" of musical elements; that is, a "vocabulary" or musical lexicon that encodes simple structures underlying the construction of larger units. Presumably this would be true for musicians/experts, but even nonmusicians may be better at attending or dancing to music in a familiar genre, which suggests some sort of familiarity with building blocks and patterns of assemblage. One clear difference will be that the lexicon carries meaning, a role that tones or musical phrases do not have (at least not in the same way; see Seifert et al., this volume). It seems that this line of argumentation thus reveals another fundamental difference between the domains.

Two additional differences are worth noting. First, while duration can function as a distinctive feature of vowels and consonants in some languages, and is also one of the acoustic correlates of word and focus stress, its role tends to be minor compared to the primary role that duration plays in musical systems (for a discussion on how to distinguish prosody from the tones of the vowels of a tone language like Chinese, see Ladd, this volume). Second, whereas the levels of linguistic structure exist, roughly speaking, on successively larger temporal frames (this is somewhat less true for signed languages), the objects described within each musical category frequently exist, to a greater degree than language, on the same timescale. Thus, meter coexists with grouping, as can be seen with respect to a minimal group of four events that occur during the metrical frame of two strong beats: a "satz" unit (roughly, melodic unit) denoted by a cadence will tend to exist on a larger timescale, often four or eight beats, etc. Furthermore, pitch relationships are experienced on a variety of temporal levels, from the most local (the motion of adjacent pitches tend to be, in most styles, primarily stepwise contours) to listeners being highly attuned to the beginnings and endings of large musical groups. Similarly, harmonic syntax is also highly locally constrained by a limited repertoire of possible progressions, which, within the so-called common practice period, achieve closure by means of the cadential progression: dominant to tonic (V–I).

Thus, we conclude that an inventory of the fundamental representational elements, as sketched out in Table 17.1, reveals a domain-specific organization, especially at the highest level of analysis. As we turn to neural evidence in sections below, the claim of domain specificity is supported by dissociations between the domains observed in both imaging and lesion data. It goes without saying, however, that there must be at least some shared attributes, to which we turn next.

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.

Domain-General, Generic Computations/Operations: Shared Attributes

In contrast to the representational inventories, we hypothesize that many of the algorithms/operations that have such primitives as their inputs are, by and large, domain general or, at least, will prove to combine generic algorithms in domain-specific ways. One way to conceptualize this is to imagine different invocations of the same neural circuitry; that is, "copies" of the same circuitry, but which operate on input representations of different types that are domain specific. For example, the task of constructing an auditory stream, of extracting relative pitch, or of sequencing or concatenating information are the types of operations that are likely to be "generic," and thus a potentially shared computational resource for processing both music and language. Some of the hypothesized shared operations are summarized in Table 17.1. (Note: we use "stream" in two senses in this chapter. We distinguish "auditory streams" as defined, for example, by the voices of different people at a cocktail party from "neuroanatomical streams" as in the two routes from primary auditory cortex to prefrontal cortex shown in Figure 17.1.) Here we offer a brief list that merits further exploration as candidate domain-general operations and then discuss two of these operations—sequencing and timing—in a bit more detail.

- *Constructing spectro-temporal auditory objects* (Griffiths and Warren 2004; Zatorre et al. 2004; Leaver and Rauschecker 2010) or *identifying auditory streams* (Shamma et al. 2011; Micheyl et al. 2005, 2007) is part of a necessary prior auditory scene analysis (Bregman 1990). The required neural circuitry is evident across species (certainly primates and vocal learners; see Fitch and Jarvis, this volume). What appears as specialization in the human brain thus arises from the interface of these more generic circuits with domain-specific input representations and/or the production and interpretation of such representations. Some of

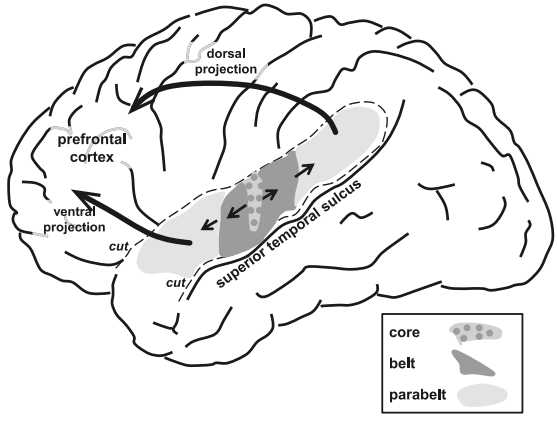

Figure 17.1 Human auditory cortex (modified from Hall and Barker 2012).

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 grouping requires attention; however some scene segregation and object identification appears to occur preattentively.

- *Discretization into auditory events*: Because both speech and music typically come to the listener as a continuous stream, some form of chunking or discretization is required (e.g., for lexical access or the identification of a theme/motif).
- *Grouping*, both in terms of establishing constituency (segmentation into groups) and hierarchy (establishing relationships between components), can occur in space (as in dance, orchestras, marching bands, choruses, and cocktail parties), in time (e.g., intricate polyrhythms in African drumming), and/or feature space (e.g., timbre). In music, lower-level components emerge at the metrical level of periodicities and metrical accents (beat induction). The next level might be harmonic. A higher level arises from the grouping of constituents and may lead to musical phrases, progressions, or the satz. In language there is a similar hierarchical process, such that you go from phonology to morphology to syntax and meaning. Building on Lerdahl and Jackendoff (1983), Drake (1998) summarizes such segmentation (as well as regularity extraction) processes and distinguishes between those that appear to be universal and innate (segmentation, temporal regularity extraction) from those that are acquired or derived and culture specific. Later (see section on Learning and Development), the way in which "innate" skills emerge over the first few months or years of life is characterized.

The above three types of operations are functions of the auditory system, and probably the cortical auditory system. The nuts and bolts of these operations are the focus of much current auditory neuroscience research (Schnupp et al. 2010). We know relatively little about the precise locations and mechanisms that are involved, beyond the rather superficial insight that superior temporal areas of the auditory cortex are implicated, including the so-called core, belt, and parabelt regions. Figure 17.1 (modified from Hall and Barker 2012) illustrates the anatomy of the human auditory cortex. It provides the gyral and sulcal anatomic context, highlights the structure of the superior temporal gyrus (STG; the region above the superior temporal sulcus), and shows one of the dorsal and ventral projection schemes with the ventral route serving as a "what" pathway associated with auditory object identification (Rauschecker and Scott 2009). To date, however, there is no compelling evidence to suggest that any of these early cortical regions are selectively specialized for either speech or music processing.

- *Sequencing of constituents* must be accomplished in various contexts involving relative time, ordering, concatenation, and relative pitch contour. Placing items (elementary representations) in a sequence must be done in both domains, but clearly differs between music and language. In language, concatenation and linear order are not sufficient—certainly

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.

for syntax and semantics, where structure dictates interpretation—although it is a critical part of phonological processing. Sequencing and ordering operations have implicated auditory areas as well as inferior frontal and premotor and motor regions, including Broca's region, basal ganglia, the cerebellum, and other potential substrates. One cortical region that is consistently implicated in basic constituent building, at least in language, is the left anterior temporal lobe (see Figure 17.1). For example, recent work by Bemis and Pylkkänen (2011) shows how minimal unit building (e.g., "red boat") activates the left anterior temporal lobe across studies and imaging approaches.

- *Linking multimodal objects* in audiovisual speech (i.e., vision and hearing), song (i.e., words/speech and melody/music), and dance (i.e. music and motoric patterns). Typically, multimodal perceptuo-motor tasks have implicated three regions: posterior superior temporal sulcus (STS; Figure 17.1), the inferior parietal lobe, and inferior frontal regions, together often referred to as the dorsal auditory cortical pathway (Rauschecker and Scott 2009). Although this has been an active area of research in multisensory speech, less is known about the musical case. What is known about song and dance is reviewed by Janata and Parsons (this volume).
- *Coordinate transformations are ubiquitous* (e.g., from input spoken words in acoustic coordinates to output speech in motor coordinates; the mapping from auditory input to vocal tract output in vocal learners; the alignment between perception and action in music performance; the alignment between musical and linguistic information in song; the alignment between musical and motor information in dance). The common problem is that information in one domain, represented on coordinate system *i*, must be made compatible with information in another domain, represented in coordinate system *j*. In the case of speech, a growing literature suggests that the dorsal processing stream (Rauschecker and Scott 2009), or perhaps more specifically a temporo-parietal area (Sylvian parieto-temporal, SPT) provides the cortical substrate for this computation (Hickok et al. 2003; Hickok and Poeppel 2007; Hickok 2012). It is not obvious where and how such basic and widespread operations are executed in tasks involving music perception and performance.
- *Entrainment, synchronization, alternation, interleaving, turn-taking* require that the listener form a model of a conversational (or musical) partner as well as an accurate internal model in order to synchronize all information and set up the framework for properly timed communicative alternation (Levinson 1997, this volume). The neural basis for these operations is not yet understood. While entrainment to stimulus features (e.g., the temporal structure of speech or music) is known to occur in sensory areas and has been well described and characterized

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.

(e.g., Schroeder et al. 2008), how such inter-agent alignment occurs in neural terms is not yet clear. There are occasional appeals to the mirror neuron system, but many in our working group showed little enthusiasm for mirror neurons and their promise (for a critique of mirror neuron hypothesis, see Hickok 2008; Rogalsky et al. 2011a; for a more positive view, see Fogassi as well as Arbib and Iriki, this volume, and Jeannerod 2005). However, observations of auditory responses in motor areas and motor responses in auditory areas highlight the importance of audio-motor linkages and transformations.

- *Organizing structure for use by social partners* (of particular relevance in conversation, musical ensemble playing and jazz improvisation, and in dance).

Sequencing in Music and Language: The Importance of Relative Pitch and Duration

The acoustic signals and production gestures of speech and music are physically complex and continuous. In both domains, a process of discretization in auditory cortex yields elementary units (such as tones in music or syllables in language) that serve as "elementary particles" for sequencing operations. These sequencing operations encode the order and timing of events, and also concatenate elementary events into larger chunks. Important work on the cognitive neuroscience of sequencing has been done by Janata and Grafton (2003), Dominey et al. (2009), and others; see also the "events-in-time" modeling described by Arbib et al (this volume).

While sequencing necessarily involves the encoding of "absolute features" of events (e.g., duration, frequency structure), a very important aspect of music and speech processing is the parallel encoding of these same physical features in relative terms. For example, when processing a musical melody or a spoken intonation contour, we extract not only a sequence of pitches but also a sequence of relative pitches (the sequence of ups and downs between individual pitches, independent of absolute frequency). This is what allows us to recognize the same melody or intonation contour (such as a "question" contour with a rise at the end) at different absolute frequency levels. Relative pitch seems to be "easy" for humans but not for other species. Birds, for example, show remarkable ability in absolute pitch (Weisman et al. 1998, 2010) but struggle with relative pitch (e.g., Bregman et al. 2012). Monkeys, however, almost totally lack relative pitch ability (Wright et al. 2000), but extensive training can lead to a limited form of relative pitch in ferrets (Yin et al. 2010) and monkeys (Brosch et al. 2004). The human ability to perceive relative pitch readily may mark a crucial step in the evolution of language and music.

Similarly, in sequencing we encode not only absolute duration of events but also the relative durations of successive events or onsets between events (e.g., inter-onset intervals). There are similar constraints for compression 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.

dilation of time in speech and music. Humans are able to recognize immediately a tune, within rather stringent limits, when it is slowed down or sped up (Warren et al. 1991), because the pattern of relative durations is preserved, even though the duration of every element in the sequence has changed. Similar limits exist for the perception of speech when the speed of its delivery is decreased or accelerated (Ahissar et al. 2001). Similarly, sensitivity to relative duration in speech allows us to be sensitive to prosodic phenomena (such as phrase-final lengthening or stress contrasts between syllables) across changes in speech rate or the overall emphasis with which speech is produced. Thus when we think of acoustic sequences in music and language that need to be decoded by the listener, it is important to remember that from the brain's perspective a sequence of events is a multidimensional object or stream unfolding in time; that is, a sequence of absolute and relative attributes, with relative pitch and relative duration being the minimum set of relative attributes that are likely to be encoded.

Neural Basis of Timing and Attention in Time

Obviously, timing is critical for music and language. Both spoken language and music are, typically, extended signals with a principled structure in the time domain. A temporally evolving dynamic signal requires that the listener (or producer) can accurately analyze the information, parse it into chunks of the appropriate temporal granularity, and decode the information in the temporal windows that are generated. There are similar timescales for both language and music processing, presumably the consequence of basic neuronal time constants. Both domains require timing at the tens-of-millisecond timescale (e.g., analysis of certain phonetic features, analysis of brief notes and spaces between notes), a scale around 150–300 ms (associated with, e.g., syllabic parsing in speech), and longer scales relating to intonation contour. It is interesting to note the extent of overlap between the timescales used for elementary analytic processes in speech and music perception. Expressed in terms of modulation rate, the typical phenomena that a listener must analyze to generate an interpretable musical experience (e.g., beat, tonal induction, and melody recognition as well as phonemic and lexical analysis) range between approximately 0.5 and 10 Hz (Farbood et al. 2013). While there are faster and slower perceptual phenomena, these are roughly the temporal rates over which listeners perform optimally; that is, these rates constitute the "temporal sweet spot" both for music and for speech.

A fair amount of recent research has focused on how to represent and analyze temporal signals at a neural level. One approach emphasizes neural timekeepers in a supramodal timing network that includes cerebellum, basal ganglia, premotor and supplementary motor areas, and prefrontal cortex (Nagarajan et al. 1998; Teki et al. 2011). In contrast, another approach emphasizes the potential utility of neuronal oscillations as mechanisms to parse

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.

extended signals into units of the appropriate size. These oscillations are population effects which provide a framework for the individual activation of, and interaction between, the multitude of neurons within the population. The basic intuition is that there are intrinsic neuronal oscillations (evident in auditory areas) at the delta (1–3 Hz), theta (4–8 Hz), and low gamma (30–50 Hz) frequencies that can interact with input signals in a manner that structures signals (phase resetting) and discretizes them (sampling); this may provide (by virtue of the oscillations) a predictive context (active sensing). Figure 17.2 illustrates the neural oscillation hypothesis (for which experimental support is still controversial and provisional). For the domain of speech processing, the neuronal oscillation approach is reviewed in Giraud and Poeppel (2012), Zion Golumbic et al. (2012), and Peelle et al. (2012). How neuronal oscillations may play a role in speech perception is also briefly summarized by Hagoort and Poeppel (this volume). One challenge is to understand how neuronal oscillations may facilitate processing on these timescales, since they are also evident in typical musical signals. Electroencephalography (EEG) or magnetoencephalography (MEG) studies using musical signals with temporally manipulated structure will have to be employed while testing where and how music-related temporal structure interacts with neuronal oscillations. Recent MEG data show that intrinsic neuronal oscillations in the theta band are facilitated in left temporal cortex when the input is intelligible speech (Peelle et al. 2012). Future studies will need to explore whether these responses might be amplified in right temporal cortex when presented with musical signals. Using neurophysiological data from single neurons, EEG, and MEG as well as neuroimaging tools, we can explore mechanistic hypotheses about how neural responses might encode complex musical or linguistic signals and guide attention allocation.

Thus one overarching question is whether the same timing mechanisms are used in both domains. A different way of approaching the question is to ask: If a subject is trained in a perceptual learning paradigm on an interval using a pure-tone duration, will this timing information be equally available for timing a sound (musical note or syllable) in a musical or linguistic context? Evidence from behavioral studies (Wright et al. 1997; Wright and Zhang 2009) indicates that temporal interval discrimination generalizes to untrained markers of the interval, but not to untrained intervals. There is also evidence that training on temporal interval discrimination (two tone pips separated by an interval) generalizes (a) to duration discrimination of the same overall duration, (b) to motor tapping (for the trained duration only), and (c) from training in the somatosensory system to the auditory system (for the trained duration only). This insight may help us understand the supramodal timing representations that underlie language as well as music and dance performance and perception. However, it leaves open whether the transfer implicates a single shared brain system or coupling between domain-specific systems (see Patel, this volume, for further discussion of shared resources). Finally, native speakers of languages that use vowel duration as a phonetic cue have better naive performance on temporal

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.

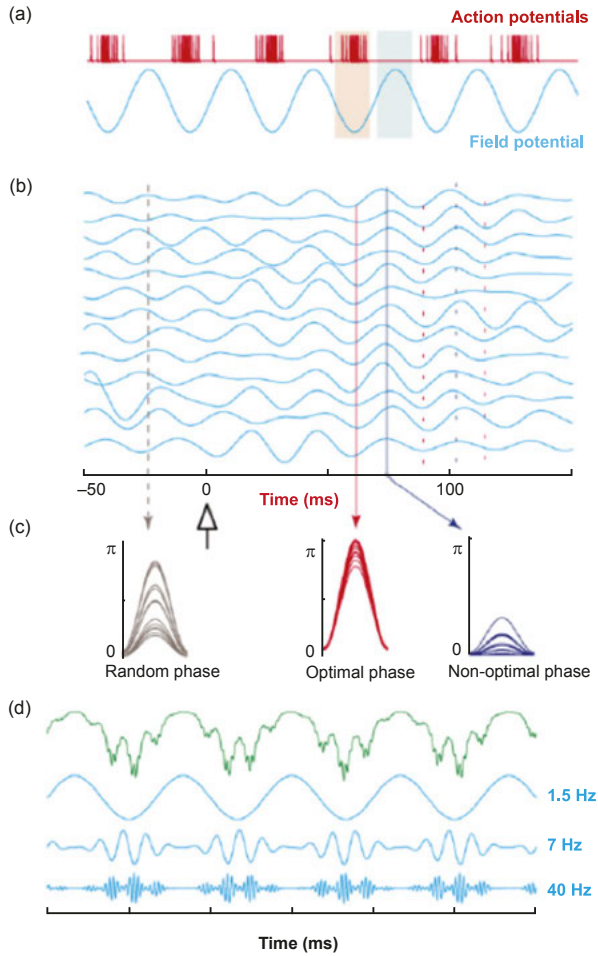

Figure 17.2 Neural oscillation hypothesis (Schroeder et al. 2008). Cortical and subcortical brain regions reveal intrinsic oscillatory neural activity on different rates/time scales, for example, between 1 Hz (delta band) and 40 Hz (gamma band). Such intrinsic oscillations are potentially in nested hierarchical relationships (d). Because the temporal structure of speech and music falls into the modulation rates of such oscillations, one hypothesis suggests that oscillatory neural mechanisms may underlie segmentation, grouping, alignment, attention allocation, and so on, and may interact with the stimulus input to generate different forms of "readout" on different timescales commensurate with the oscillations. The relation between firing patterns and excitability cycles provided by oscillations is shown in (a); the relevance of phase is depicted in (b) and (c). Intuitively, these mechanisms allow the system to lock to the phase of (or entrain to) the temporal structure of a stimulus and generate temporal windows or units for further processing. The alignment of spikes with preferential phases of a cycle (a) illustrates the packaging of spikes by oscillations. For a detailed discussion of the relevant cellular circuitry specifically for speech processing, see Giraud and Poeppel (2012). Figure from Schroeder et al. (2008), used with permission from Elsevier.

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.

interval discrimination than do native speakers of other languages, and musicians have better naive performance on temporal interval discrimination than nonmusicians, thus suggesting that timing is truly a domain-general capacity.

Cross-Cutting Approaches and Sources of Evidence

In this section, we touch on four areas of neurocognitive research that have provided data both on domain-specific representational and domain-general computational questions. We briefly discuss, in turn, input codes, learning and development, brain injury and plasticity, as well as the interaction between perception, action, and prediction.

The Role of Input Codes or How Input Can Determine Functional Specialization

The input code (in the afferent auditory pathway, and especially cortex) for music and spoken language is arguably the same kind of spectro-temporal representation (Griffiths et al. 1998) and is processed in parallel by distinct networks tuned to different features; for example, spectral versus temporal resolution (Zatorre et al. 2004). Trivially, at the most peripheral level, the signal that the system receives is the same: spectro-temporal variation stimulates the auditory periphery. Thus, differences between the domains must arise at more central levels. A fundamental issue concerns whether the structure of the input interacts with neuronal specializations of a certain type, such as preferences for spectral information versus temporal information or preferences for certain time constants.

Input codes may transform general-purpose auditory mechanisms into specialized ones that ultimately interact with the representations underlying music or speech. The existence of multiple specialized microsystems, even if they function in a similar way is more likely because modularization is more efficient. It is possible that domain specificity emerges from the operation of a general mechanism. However, in practice, it may be very difficult to demonstrate it because the general or "shared" mechanisms under study are likely to modularize with experience and also because dual domain-specific mechanisms may work together, as in song learning (Thiessen and Saffran 2003).

For example, the acquisition of tonal knowledge uses general principles by extracting statistical regularities in the environment (Krumhansl 1990; Tillmann et al. 2000). Although tonal encoding of pitch is specific to music, it may be built on "listeners' sensitivity to pitch distribution, [which is] an instance of general perceptual strategies to exploit regularities in the physical world" (Oram and Cuddy 1995:114). Thus, the input and output of the statistical computation may be domain specific while the underlying learning

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.

mechanism is replicated across circuitry serving both domains (Peretz 2006). Once acquired, the functioning of the system—say the tonal encoding of pitch—may be modular, by encoding musical pitch in terms of keys exclusively and automatically.

The same reasoning applies to auditory scene analysis as well as to auditory grouping. The fact that these two processing classes organize incoming sounds according to general Gestalt principles, such as pitch proximity, does not mean that their functioning is general purpose and mediated by a single processing system. They need not be. For instance, it would be very surprising if visual and auditory scene analyses were mediated by the same system; both types of analyses obey Gestalt principles. It is likely that the visual and auditory input codes adjust these mechanisms to their processing needs.

A developmental perspective (see next section) may be useful in disentangling initial states from modularized end stages, in both typical and atypical developing populations. Developmental disorders offer special insight into this debate. Advocates of a "domain-general" cognitive system may search for co-occurrence of impairments in music and language (and other spheres of cognition, such as spatial cognition). Such correlations may give clues as to the nature of the processes that are shared between music and language. It may turn out that domain specificity depends on very few processing components relative to a largely shared common cognitive background. These key components must correspond to domain- and human-specific adaptations, whereas the common background is likely to be shared with animals. Developmental disorders are particularly well placed to yield insight into both parts of the debate: that which is unique to music and language, and that which is not. It follows that much can be learned by comparing impaired and spared music, language, and cognition in individuals both within and between disorders over the course of development.

Still, somewhat separable modular components may exist for speech and music processing, both at a lower auditory-processing level and a higher cognitive level. Not surprisingly, the null hypothesis (analyzed by Patel, this volume) is that speech and music have very little in common in terms of cortical cognitive processing.

Learning and Development

Infants are born unable to understand or speak a particular language; they are also unable to understand or produce music. In both cases, language and music are acquired in an orderly sequence through everyday informal interaction with people in a cultural setting.

It is now accepted that the brain has a remarkable capacity to modify its structural and functional organization throughout the life span, in response to injuries, changes in environmental input, and new behavioral challenges. This plasticity underlies normal development and maturation, skill learning,

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.

memory, recovery from injury, as well as the consequences of sensory deprivation or environmental enrichment. Skill learning offers a useful model for studying plasticity because it can be easily manipulated in an experimental setting. In particular, music making (e.g., learning to sing or play a musical instrument) is an activity that typically begins early in life, while the brain has greatest plasticity. Often, musical learning continues throughout life (e.g., in musicians). Recent high-resolution imaging studies have demonstrated the ability of functional and structural auditory-motor networks to change and adapt in response to sensorimotor learning (Zatorre et al. 2012b).

Returning to the Marr-inspired taxonomy of levels, the representational elements of language and music are different, as shown in Figure 17.1. Those of language include phonemes, morphemes, words, and phrases whereas those of music include notes, pitch intervals, beats and meters, motifs and melodies. Despite representational differences at the higher level, music and language do rely to a large extent on shared elementary procedures that appear to be in place in prelinguistic and premusical infants (Drake 1998; Trainor and Corrigal 2010).

Indeed, it is possible that domain-specific processing develops in the brain largely through exposure to the different structures in the speech and musical input from the environment. Of course, at the same time, specialization of some brain regions for music or language likely occurs from intrinsic properties of those regions being more suited for processing structural elements of music (e.g., fine spectral structure) or speech (e.g., rapid temporal structure). Receptive language and music tend to be processed in similar regions in most people, though with some hemispheric differences, and expressive speech and musical vocalization might rely on shared auditory-motor systems (Özdemir et al. 2006).

To demonstrate the capabilities of infants (which are not, in most cases, present in neonates), consider the following examples of the early presence of a number of putatively primitive domain-general processing mechanisms (for more on this topic, see Trehub):

1. *Constructing spatiotemporal objects*: Newborns are able to discriminate their mother's voice from that of a stranger (DeCasper and Fifer 1980). At 6 months (probably younger), they can discriminate one voice from another in the context of multiple tokens from each speaker. That there is a learned component to this is evident: infants are equally good at human and monkey voice discrimination at 6 months, but better for human voices at 12 months. Infants can also discriminate timbres, and exposure to a particular timbre increases their neural response to that timbre, as measured by EEG (Friendly, Rendall, and Trainor, pers. comm.). From at least as young as two months, infants can categorize musical intervals as consonant or dissonant and prefer to listen to consonance (Trainor et al. 2002).

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.

2. *Discretizing and sequencing* the signal: Young infants can discriminate rhythmic patterns as well as orders of pitches in a sequence (Chang and Trehub 1977; Demany et al. 1977)
3. *Relative pitch:* Infants readily recognize melodies in transposition, as evidenced by behavioral and EEG (mismatch response) studies (e.g., Tew et al. 2009; Trainor and Trehub 1992).
4. *Relative timing*: Infants recognize sequences played at somewhat faster or slower rates (Trehub and Thorpe 1989).
5. *Grouping*: Infants segregate and integrate incoming elements into perceptual streams. This has been shown both for sequential input, where higher and lower tones are grouped into separate streams (Winkler et al. 2003) as well as for simultaneous input, where one of several simultaneous tones can be "captured" into a separate stream if the simultaneous tones are preceded by several repetitions of the tone to be captured (Smith and Trainor 2011). Similarly, infants hear a harmonic that is mistuned as a separate object in a complex tone (Folland et al. 2012). The presentation of repeating patterns (e.g., short–short–long) also leads to grouping, such that group boundaries are received after the "long" elements.
6. *Hierarchical processing*: Infants perceive different meters, which require processing on at least two levels of a metrical hierarchy (Hannon and Trehub 2005a).
7. *Coordinate transformations and sensorimotor coordination*: Because young infants are not motorically mature, this is more difficult to demonstrate. The way that they are moved by their caregivers, however, affects their perception of auditory patterns, suggesting that they can transform from one reference frame to another (Phillips-Silver and Trainor 2005). When infants are presented with a repeating auditory six-beat pattern with an ambiguous meter (i.e., it had no internal accents), the pattern can be interpreted as two groups of three beats (as in a waltz) or as three groups of two beats. During a training phase, two groups of infants heard the ambiguous rhythm while they were simultaneously bounced up and down: one group on every third beat, the other on every second beat of the pattern. After this familiarization, both groups were given a preferential listening test. Infants bounced every second beat of the ambiguous pattern preferred (in the absence of bouncing) to listen to a version with accents added every second beat compared to a version with accents added every third beat. On the other hand, infants bounced on every third beat of the ambiguous pattern preferred to listen to the version with accents every third beat compared to the version with accents very second beat. The fact that infants were passively bounced and that the effect remained when they were blindfolded suggests that the vestibular system may play a role in

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. This study also suggests that the roots of common representations for music and dance may be seen in infancy.

8. *Prediction*: Auditory mismatch responses in EEG data (mismatch negativity) can be seen very early in development, even *in utero* during the last trimester (Draganova et al. 2005). In these studies, one stimulus or set of stimuli was repeated throughout; occasionally, one repetition was replaced with another, deviant stimulus. The existence of a mismatch response suggests that sensory memory is intact and that the mechanisms underlying regularity extraction and local prediction in time are available at the earliest stages of development.
9. *Entrainment/turn-taking*: Any evidence for entrainment in young infants is not widely known, although how they are moved to rhythms by their caregivers affects how they hear their metrical structure (Phillips-Silver and Trainor 2005). However, there is evidence for turn-taking in speech interactions with adults.

This short review suggests that the basic processing algorithms that enable language and musical learning in the young infant are in place as of a very early age. However, it goes without saying that the "linguistic and musical inventory" is not yet in place at this stage. That is, the representational elements are acquired, incrementally, over the course of development. In the case of speech and language, the trajectory is well known. In the first year, the learner acquires the sounds (signs) of her language, and by the end of year 1, the first single words are evident. Between the ages of two to three years, the vocabulary explosion "fills" the lexicon with items, and the first structured multiword (or multisign) utterances are generated. There is consensus that by three years of age, the neurotypical learner has the syntactic capabilities of a typical speaker (with, obviously, a more restricted vocabulary). What the steps look like for a child learning music—perhaps through song—is less clear.

The Give and Take of Language and Music

The Perception–Prediction–Action Cycle

The perception–action cycle (Neisser 1976; Arbib 1989) emphasizes that we are not bound by stimuli in our actions. In general, our perceptions are directed by our ongoing plans and intentions, though what we perceive will in turn affect our plans and actions. Within this framework, Fuster (2004) postulates that (a) action plans are hierarchically organized in the frontal lobe (Koechlin and Jubault 2006) whereas perception is hierarchically organized in the temporal, occipital, and parietal lobes, and (b) reciprocal paths link action and perception at all levels. One may recall the work of Goldman-Rakic (1991) in delineating the reciprocal connections between specific areas of frontal and parietal cortex.

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.

Janata and Parsons (this volume) discuss this further and emphasize that attentive listening to music can engage the action systems of the brain.

A key part of the perception–action cycle is the predictive model: to prepare the next action, it is important to generate plausible expectations about the next stimulus. Activity in the auditory cortex thus represents not only the acoustic structure of a given attended sound, and other sounds in the environmental soundscape, but it also signals the predicted acoustic trajectories and their associated behavioral meaning. The auditory cortex is therefore a plastic encoder of sound properties and their behavioral significance.

Feedback and predictive (feedforward) coding is likely to function in both the dorsal and ventral auditory streams (Rauschecker and Scott 2009; Rauschecker 2011; Hickok 2012), with the direction of feedback and feedforward depending on one's vantage point within the perception–action cycle. During feedback from motor to sensory structures, an efference copy sent from prefrontal and premotor cortex (dorsal stream) could provide the basis for sensorimotor control and integration as well as "optimal state estimation" in the inferior parietal lobe and in sensory areas of the posterior auditory cortex. In contrast, forward prediction may arise from the ventral stream through an "object-based" lexical–conceptual system.

Figure 17.3 complements Figure 17.1 to provide a perspective on the interaction between auditory, premotor, and prefrontal areas using the notion of internal models. The perception–action cycle can be run either as a forward or an inverse model. The predictive, forward mapping builds on knowledge about objects processed and stored in the anterior temporal lobe via the ventral stream and continues via prefrontal and premotor cortex into parietal and posterior auditory cortex, where an error signal is generated between real and predicted input. The inverse mapping, which runs the cycle in the opposite direction, instructs the motor system and creates affordances via the dorsal stream for generating sounds that match the motor representations, including sound sequences that require concatenation in a particular order, as they are the substance of both speech and music. There is overwhelming evidence for such internal forward models in motor control (Flanagan et al. 2003; Wolpert et al. 2003; Wolpert and Kawato 1998), but the extension to both perceptual and cognitive models is more recent.

How much prediction occurs at this level of neuronal precision in the human auditory cortex as we process speech? One of the domains in which this has been addressed extensively is audiovisual speech. Both EEG and MEG research (e.g., van Wassenhove et al. 2005; Arnal et al. 2009) and fMRI-based (Skipper et al. 2007a, b, 2009) research has convincingly shown that information conveyed by facial cues provides highly predictive and specific information about upcoming auditory signals. For example, because facial dynamics slightly precede acoustic output (Chandrasekaran and Ghazanfar 2009), the content of the face signal (e.g., bilabial lip closure position) signals that a certain consonant type is coming (e.g., "b," "p," or "m"). This prediction is

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.

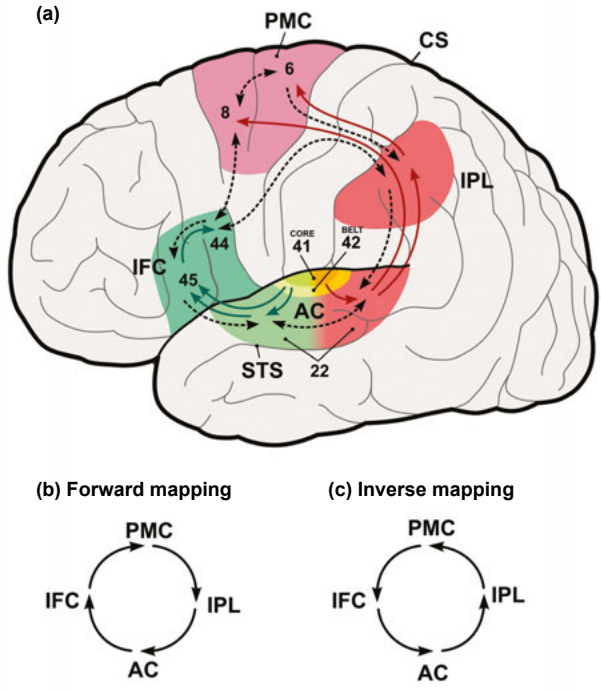

Figure 17.3 Feedforward and feedback organization (after Rauschecker and Scott 2009). (a) A schematic version of the dorsal and ventral processing streams and their basic connectivity. Dorsal projections extend from superior posterior temporal (auditory cortex, AC) through inferior parietal lobe (IPL) to inferior frontal cortex (IFC) and premotor cortex (PMC). The ventral stream projections typically extend through the extreme capsule and the uncinate fasciculus to inferior frontal areas. Superior temporal sulcus: STS; CS: central sulcus.The hypothesized forward and inverse mappings are illustrated in (b) and (c), respectively.

reflected both in response latencies (shorter for highly predictable items) and amplitudes (smaller for predictable items). Overall, whether one is considering speech or music alone as purely acoustic signals, audiovisual signals, such as speaking faces or playing musicians, or spoken sentences, or even higher-order conceptual information, it is beyond dispute that both high- and low-level information is incorporated into models (perhaps of a Bayesian flavor) that shape upcoming perception and action in a precise manner. There is as much, if not more, expectancy-driven top-down processing as there is bottom-up analysis. Of course, prediction occurs at higher levels, as when the listener predicts the next word of a sentence.

Feedback is also critical at lower levels. Speech production is known to be dependent on auditory feedback, going back to Levelt (1983) and emphasized in recent work by Houde and Nagarajan (2011) and Hickok et al. (2011).

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.

Speech can lead to motor-induced suppression of the auditory cortex (Allu et al. 2009) and may result in noise suppression or cancellation of self-produced speech. MEG studies have long shown evidence for this putative efference copy, i.e., a predictive signal from motor to auditory cortex (Kauramäki et al. 2010; Nishitani and Hari 2002), and nonhuman primate studies have demonstrated a neurophysiological correlate (Eliades and Wang 2008). One insight that has emerged is that local, early feedforward as well as feedback processing in auditory cortical areas reflects analysis of the error signal (i.e., the mismatch between the predicted input and the actual input).

Recent research has been conducted on brain activation during total silence, based on the expectation of upcoming or anticipated sounds. For a recent example of this in musical sequences, see Leaver et al. (2009).

Timing and Turn-Taking

Elsewhere in this volume, Levinson explores "the interactive niche," which includes social interactions involved in turn-taking as well as the sequencing of actions. "Informal verbal interaction is the core matrix for human social life. A mechanism for coordinating this basic mode of interaction is a system of turn-taking that regulates who is to speak and when" (Stivers et al. 2009:10,587). The work of Levinson and his colleagues (Stivers et al. 2009) has shown that there are striking universals in the underlying pattern of response latency in conversation, providing clear evidence for a general avoidance of overlapping talk and a minimization of silence between conversational turns (an incredibly brief gap since the peak of response is within 200 ms of the end of the previous question). As Levinson observes, since it takes at least 600 ms to initiate speech production, speakers must anticipate the last words of their companion's turns and predict the content and the form of their companion's utterance in order to respond appropriately. Thus, conversation is built on detailed prediction: figuring out when others are going to speak, what they are going to say, when they are going to finish, and how to prepare your own reply. This requires encoding of the utterance they intend to make at all levels.

Another study on turn-taking (De Ruiter et al. 2006) demonstrates that the symbolic (i.e., lexical, syntactic) content of an utterance is necessary (and possibly sufficient) for projecting the moment of its completion, and thus for regulating conversational turn-taking. By contrast, and perhaps surprisingly, intonational contour is neither necessary nor sufficient for end-of-turn projection. This overlap of comprehension and production in conversation can be extremely demanding at a cognitive level.

As Hagoort and Poeppel (this volume) state:

> It might well be that the interconnectedness of the cognitive and neural architectures for language comprehension and production enables the production system to participate in generating internal predictions while in the business of

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.

comprehending linguistic input. This prediction-is-production account, however, might not be as easy in relation to the perception of music, at least for instrumental music. With few exceptions, all of humankind are expert speakers. However, for music, there seems to be a stronger asymmetry between perception and production. Two questions result: Does prediction play an equally strong role in language comprehension and the perception of music? If so, what might generate the predictions in music perception?

Clearly, predictions can be guided if the musicians are playing a composed score of music from memory. Thus, this question is likely to be particularly important during conversational turn-taking in music, which may place even greater cognitive demands on a musician playing in an orchestra or quartet, and particularly during improvisation, as in jazz. Moreover, playing in an ensemble, singing in a choir or dancing with a partner all involve patterns of coordination that require far more delicate timing than that involved in initiating a turn in a conversation.

The true complexity of the mechanics of turn-taking is illustrated in Figure 17.4 (Menenti et al. 2012). Building on the work of Pickering and Garrod (2004), Figure 17.4 shows at how many levels of analysis two speakers have to align, ranging from sounds to highly abstract situation models. In the case of language, the nature of the necessary alignments becomes increasingly well understood. However, whether such a model is plausible (or even desirable) for musical performance or dance in a pair or group is not at all clear. Future

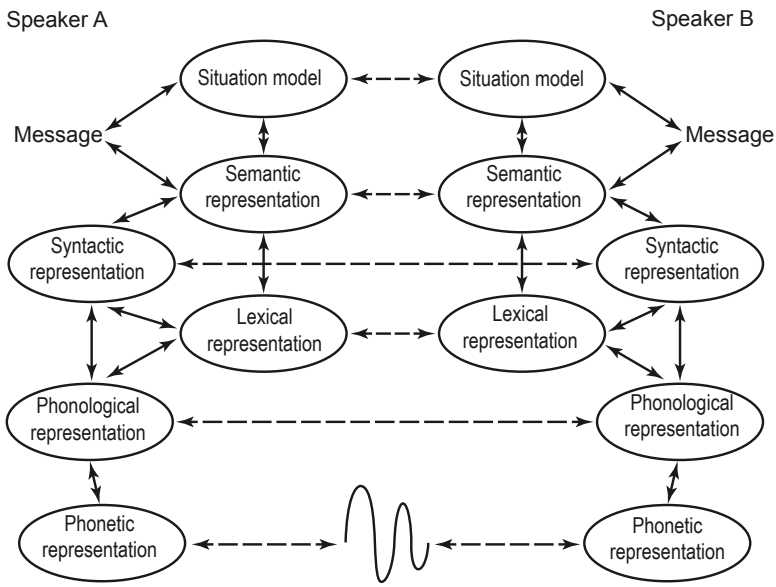

Figure 17.4 The interactive alignment model (reprinted with permission from Menenti et al. 2012).

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.

work will have to determine if such alignments are in play at all, in ways similar to conversation models.

The neural basis for the postulated "representational parity" between production and comprehension is a topic of current research (Menenti et al. 2012). To test these predictions, future studies will record from participants engaged in online interaction.

Perception and Production Interaction in Singing

Although this topic has been discussed by Janata and Parsons (this volume), we wish to emphasize the importance of the interaction between perception and production in singing. Vocal control of song and pitch has been studied in both nonmusicians and musicians (Perry et al. 1999; Brown et al. 2004; Zarate and Zatorre 2008). Recent behavioral studies report evidence that suggests that nonmusicians with good pitch discrimination sing more accurately than those with poorer auditory skills (perceptual and vocal production skills). However, Zarate et al. (2010) gave auditory discrimination training on micromelodies to a group of nonmusicians and found that training-enhanced auditory discrimination did not lead to improved vocal accuracy although it did lead to enhanced auditory perception (Zatorre et al. 2012a). Thus, there may be a partial dissociation between auditory perceptual and vocal production abilities; that is, while it may not be possible to produce precise pitch intonation without equal perceptual abilities, the presence of perceptual ability alone does not guarantee vocal ability. (For differences in the brain systems supporting verbal and tonal working memory in nonmusicians and musicians, see Koelsch, this volume.)

Audiomotor interactions in music perception and production have been reviewed by Zatorre et al. (2007). Auditory imagery has also been described by Leaver et al. (2009), suggesting links between auditory and motor areas, premotor and supplementary motor areas, as well as prefrontal regions. Zatorre (2007) proposes that when we listen to music, we may activate the ventral premotor cortex links, associated with producing the music. However, listening also engages another neural system, in which the dorsal premotor cortex is a component, to process higher-order metrical information, which may be critical for setting up temporal and melodic expectancies at the heart of musical understanding. These topics are discussed further below in the section on the neurology of amusia.

Neurobiological Constraints and Mechanisms

The neurobiological foundations of music, speech, and language processing have been studied at virtually every level of analysis: from single unit physiology to noninvasive imaging to deficit lesion studies. Here we offer a selection of phenomena to illustrate the range of data that need to be incorporated into

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.

a synthetic understanding of shared and distinctive processes in music and language.

Domain-General Processes: Shared Neural Substrates for Language and Music

Hierarchical Organization

There is considerable evidence for hierarchical organization in the human auditory cortex (Okada et al. 2010; Obleser et al. 2010; Chevillet et al. 2011; DeWitt and Rauschecker 2012) as well as in nonhuman primates (Tian et al. 2001; Rauschecker and Scott 2009; see also Figure 17.1 and 17.3). Core regions, including primary auditory cortex, respond best to tones and contours, whereas higher areas (belt and parabelt) are selectively responsive to more complex features such as chords, band-passed noise, vocalizations, and speech. We propose that core levels encode low-level features that are combined in higher levels to yield more abstract neural codes for auditory objects, including phonemes and words for language.

Although neural representation of music is likely to be constrained by this hierarchical organization for auditory processing in the brain, one key issue is whether such hierarchical organization can be demonstrated for higher groupings in music. In the case of language, we have pretty clear ideas of what constitutes a processing hierarchy (e.g., phonology, morphology, lexical semantics, syntax, compositional semantics, discourse representation), and there is a growing body of evidence about where such functions are executed (for a review, see Hagoort and Poeppel, this volume). How each level is executed, however, is largely unknown. In the case of music, there are equally intuitive hierarchies (e.g., note, motif, rhythm, melody, piece), but the functional anatomy of the hierarchy is less clear. Trivially, auditory areas are implicated throughout; interestingly motor areas are implicated in many of the temporal, beat, and rhythm subroutines. However, a well-defined functional anatomy is still under construction, in part because many of the functions are compressed into auditory regions and the role of memorized structures is less well established (see below).

Acoustic Scene Analysis and Streaming

Auditory streaming is the perceptual parsing of acoustic sequences into "streams." This makes it possible for a listener to follow sounds from a given source despite the presence of other sounds and is critical in environments that contain multiple sound sources (Carlyon 2004). Neural mechanisms underlying streaming are common to music and language, are strongly influenced by attention, and appear to use a full range of grouping mechanisms for frequency,

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.

timbre, as well as spatial and temporal cues (Micheyl et al. 2005; Shamma et al. 2011; Wang and Brown 2006).

Real-Time, Attention-Driven Adaptive Plasticity

To understand what is going on at a neural level, it is critical to realize that the auditory cortex is not a passive detector of acoustic stimulus events. Its activity and responses nimbly change with context (reflecting task demands, attention, and learning) to provide overall functional relevance of the sound to the listener (Fritz et al. 2003). Rapid changes in auditory filter properties reconfigure the listening brain to enhance the processing of current auditory objects of interest, whether in the linguistic or musical domain, and may help segregate relevant sounds from background noise (Ahveninen et al. 2011). The attention-driven top-down capabilities are especially important in light of the top-down influences (evident in the perception–action cycle) that condition processing even at the periphery (Fritz et al. 2010; Xiao and Suga 2002; Delano et al. 2007, 2010; Leon et al. 2012). Figure 17.2 illustrates one possible hypothesis of how such attention in time can be accomplished. If either musical or speech elements arrive at predictable times (as they often do, given underlying rhythms, though more so in music), amplifying or selecting those moments can facilitate processing with attention-driven, adaptive plasticity mechanisms. A challenge for this oscillatory hypothesis arises in situations where acoustic input is less structured in time.

Pitch

There are many examples of (nearly) perfect pitch and perfect tempo in musicians. What was a bit unexpected is that there is also good evidence that mothers without musical training also demonstrate absolute pitch and tempo as they sing songs to their infants (Bergeson and Trehub 2002). It is also intriguing to note that musicians who are native speakers of a tone language are more likely to have musical absolute pitch than musicians who do *not* speak a tone language (Deutsch et al. 2006).

What is the neurobiological basis of pitch? There are well-studied neurophysiological mechanisms that can help us begin to think about how absolute pitch is encoded (Bendor and Wang 2005; Bizley et al. 2009) as well as how relative pitch is encoded, such as frequency-shift detectors or frequency-modulated, direction-sensitive neurons in the auditory cortex. Computational models have built on this work to suggest how the brain represents and remembers sequences of relative pitches (see, e.g., the model put forth by Husain et al. 2004), which incorporates multiple brain regions, including superior temporal and prefrontal regions.

Interestingly, studies of other species with complex acoustic communication (such as starlings) show that recognizing tone sequences on the basis of relative

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.

pitch is difficult for nonhuman animals (cf. Bregman et al. 2012). This raises the question of whether our system has been optimized or specialized over evolutionary time for this purpose. Animal behavioral studies are beginning to address this issue (Wright et al. 2000; Yin et al. 2010; Bregman et al. 2012) as are some initial animal neurophysiological studies (Brosch et al. 2004).

Timing and Beat Perception

A large body of work, including recent neurophysiological studies (Jaramillo and Zador 2011; Bendixen et al. 2011), shows that neurons in auditory cortex are modulated by the expected time of arrival of an incoming sound. What is the neural basis for timing and time constants?

Timing networks are widespread throughout the brain. A recent experimental and theoretical study by Bernachhia et al. (2011) suggests that there is a neuronal "reservoir" of time constants in areas of the prefrontal, cingulate, and parietal cortex that can be used to support a flexible memory system in which neural subpopulations with distinct sets of long or short memory timescales can be deployed according to task demands. Other studies (Itskov et al. 2011; Jin et al. 2009; Fritz et al. 2010) have shown similar arrays of neurons with variable time constants in cerebellum, basal ganglia and prefrontal cortex.

Interestingly, simply listening to musical rhythms activates the motor system (Chen et al. 2008a). The cerebellum, basal ganglia, dorsal premotor cortex, and prefrontal cortex have all been shown to play an important role in timing in music, and likely in language processing (Zatorre 2007). A recent imaging study (Teki et al. 2011) suggests that there are distinct neural substrates for beat-based and duration-based auditory timing encompassing a network of the inferior olive and the cerebellum that acts as a precision clock to mediate absolute, duration-based timing, and a distinct network for relative, beat-based timing incorporating a striato-thalamo-cortical network. The supplementary motor area (SMA) and pre-SMA are critical for sequencing and integration into unified sequences (Bengtsson et al. 2009; Leaver et al. 2009). However, these networks are not typically recruited during language processing, notwithstanding the quasi-rhythmic nature of spoken language (see, however, Ivry et al. 2001. However, we do not yet know exactly how the concatenation of elements into specific sequences is accomplished in musical perception and production; neither do we know how such sequences are exquisitely timed, and how sequences and their tempo are recalled. One recent study that has explicitly addressed the issue of the timing circuit, especially with regard to the role of the basal ganglia, is by Kotz et al. (2009). They develop the well-known view that the basal ganglia play a key role in sequencing motor production to argue for its role in sensory predictability in auditory language perception.

Although there is insight into neural mechanisms that may underlie timing, the neural basis for beat and meter processing is still largely unknown. The resonance hypothesis for beat and meter perception (Large 2008) proposes that

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.

beat perception emerges from the entrainment of neuronal populations oscillating at the beat frequency, giving rise to higher-order resonance at subharmonics of beat frequency, corresponding to the meter. Experimental support for the resonance hypothesis comes from a recent study showing that entrainment to beat and meter creates temporal expectancies which can be observed directly in the human EEG as a periodic response at the frequency of the beat and at subharmonics corresponding to the metrical interpretation of the beat (Nozaradan et al. 2011). Although such entrainment clearly occurs in music, it is also likely to occur in poetry and cadenced speech (see also Ladinig et al. 2009; Honing et al. 2009).

Self-Monitoring During Speech and Music

Vocal communication involves both speaking and listening, often taking place concurrently. It is important for the auditory system to simultaneously monitor feedback of one's own voice as well as external sounds coming from the acoustic environment during speaking. The self-monitoring in the audio-vocal system may play a part in distinguishing between self-generated or externally generated auditory inputs and also in detecting errors, and making compensatory corrections, in our vocal production as part of a state feedback control system (Houde and Nagarajan 2011). Neurons in the auditory cortex of marmoset monkeys are sensitive to auditory feedback during vocal production, and changes in vocal feedback alter the coding properties of these neurons and increased their sensitivity (Eliades and Wang 2003, 2005, 2008). Such self-monitoring occurs during speaking and singing as well as during instrumental performance. In addition, there is clear evidence for attenuation or suppression of neural responses to self-triggered sounds in the human auditory cortex (motor-induced suppression for nonvocal as well as vocal stimuli). This suggests the importance of internal forward-predictive models in processing sound and distinguishing between the auditory consequences of one's own actions as distinct from other externally generated acoustic events (Baess et al. 2011; Martikainen et al. 2005). Musicians have been shown to have a particularly keen ability to generate accurate forward-predictive models (Tervaniemi et al. 2009). Thus, more generally, prior expectation (based on memory or forward models) can mediate neural adaptation (Todorovic et al. 2011).

Auditory Memory

Human long-term auditory memory appears to be extraordinarily powerful when it comes to recall of poetry or music, as compared to individual acoustic stimuli, even in musicians—particularly in comparison with the striking retention observed in visual memory (Cohen et al. 2009, 2011). Visual recognition memory in monkeys is much superior to auditory recognition memory, for stimuli with no visual association (Fritz et al. 2005). These results support the

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.

hypothesis that the emergence of language and music went hand in hand with the development of improved auditory working memory and long-term memory for abstract but meaningful sounds (Aboitiz et al. 2010; Aboitiz and García 2009). Neuroimaging evidence also exists for a network of at least two distinct, and highly interconnected, neural loci for working memory, which are both part of the dorsal cortical pathway (Rauschecker and Scott 2009)—a frontal region which comprises the dorsal part of Broca's area (Brodmann area 44) and the adjacent inferior frontal sulcus—supporting syntactic working memory and a phonological working memory store in parietal cortex (Friederici 2012). Learning the pathways and dynamic interactions between these two areas will greatly aid our understanding of language processing.

There may be evidence for a dissociation of memory for melody and lyrics in song. A patient with a lesion of the right hemisphere anterior temporal lobe and right hemisphere lateral prefrontal cortex (Steinke et al. 2001) was able to recognize familiar songs when the accompanying lyrics were removed (i.e., melodies without words), but could not recognize equally familiar but purely instrumental melodies. Evidence for the integration of melody and text has been found, however, in other patients, such as expressive aphasics who can accurately sing but not speak the lyrics of familiar songs.

Exemplar-based verbal memory has been shown in the linkage of identification and memory for individual voices with word recall ability (linking "who" and "what"; Nygaard and Pisoni 1998; Perrachione et al. 2011). In terms of linking melody to instrumental timbre ("which musical instrument played that piece" or motor linkages, i.e., "how I played that piece on that instrument" and "what melody was played"), this has also been demonstrated in musical memory. Such exemplar-based memory for music has even been observed in infants (Trainor et al. 2004). The linkage of auditory memory with semantics may also be inferred from patients with semantic dementia, who have difficulty in understanding the meaning of environmental sounds and are also impaired in the recognition of melodies (Hsieh et al. 2011). Even within the context of verbal material, there is better memory for poetry than for prose (Tillmann and Dowling 2007), emphasizing the mnemonic importance of temporal organization and rhythmic structure in poetry and music, and thus linking memory for words with rhythm and rhyme.

Koelsch (this volume) reviews evidence for two auditory working memory systems: a phonological loop which supports rehearsal of phonological (verbal) information, and a tonal loop which supports rehearsal of tonal (nonverbal) information, which are differentially developed and localized in musicians and nonmusicians. Furthermore, there are different short-term, or sensory-memory storage buffers for pitch, timbre, loudness, and duration (Semal and Demany 1991; Clement et al. 1999; Jaramillo et al. 2000; Caclin et al. 2006) and different working memory networks for melodies and rhythms (Jerde et al. 2011). Studies also indicate that auditory short-term memory for complex tone patterns is enhanced in musicians (Boh et al. 2011).

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.

Domain-Specific Processes: Neural Substrates for Music

Deficit-Lesion Characterizations: Insights from Neurological Cases

Many neuropsychological dissociations exist between language and music perception and production (Peretz 2006). One very rare but compelling neurological disorder is pure word deafness. There have been only a few dozen documented cases since the late 1800s, and very few cases in which the syndrome is "pure," but there is convergence on the general phenomenon (for reviews, see Poeppel 2001; Stefanatos 2008). Such patients have normal audiograms and largely intact peripheral auditory processing. Moreover, they are not aphasic; that is, they can read, write, and *speak*, albeit often haltingly. However, their perception of spoken language, and indeed any speech stimulus, is completely compromised. Patients are deaf to spoken words, yet, interestingly, their perception of music is relatively intact. Thus, in one single dissociation, hearing, speech, perception, language, and music are functionally fractionated. Recent cases (Stefanatos 2008; Wolmetz et al. 2011; Slevc et al. 2011) support the conjecture that the lesion pattern underlying pure word deafness is twofold. In two-thirds of the cases, the posterior aspect of the STG is affected, bilaterally. (This means patients have two lesions, sequentially.) In one-third of the cases, a deep left lateralized white matter lesion is observed; this lesion deafferents the two sides from one another, thus also implicating the integrity (or integration) of both sides for successful speech processing. Posterior STG/STS has long been thought of as the necessary tissue for speech perception, but not music perception. A recent meta-analysis of a large number of neuroimaging studies of speech perception argues, however, that the necessary site is, in fact, more anterior than previously thought (DeWitt and Rauschecker 2012). Thus, the relative contributions of each site are still under debate.

Amusia and Congenital Amusics

Parallel to pure word deafness, a well-characterized neuropsychological disorder is acquired amusia, where patients have selective difficulty with processing musical material. Brain lesions can selectively interfere with musical abilities while the rest of the cognitive system remains essentially intact (e.g. Steinke et al. 1997). Conversely, brain damage can impair musical abilities exclusively. Patients may no longer recognize melodies (presented without words) that were highly familiar to them prior to the onset of their brain damage but perform normally when recognizing spoken lyrics (and words, in general), familiar voices, and other environmental sounds (e.g., animal cries, traffic noises, and human vocal sounds). The existence of a specific problem with music alongside normal functioning of other auditory abilities, including speech comprehension, is consistent with damage to processing components

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 are both essential to the normal process of music recognition and specific to the musical domain (for reviews, see Peretz 2006, Peretz et al. 2009a).

Similar findings are obtained in production studies. Brain-damaged patients may lose the ability to sing familiar songs but retain the ability to recite the lyrics and speak with normal prosody. The reverse condition (i.e., impaired speech with intact vocal production) is more common or has been reported more often. Aphasic patients may remain able to sing familiar tunes and learn novel tunes but fail to produce intelligible lyrics in both singing and speaking (Racette et al. 2006). These results suggest that verbal production, whether sung or spoken, is mediated by the same (impaired) language output system, and that this speech route is distinct from both the (spared) musical and prosodic route. In sum, the autonomy of music and language processing extends to production tasks.

Similarly, individuals who suffer from lifelong musical difficulties, a condition which Peretz (2008) and Stewart (2011) refer as to congenital amusia, have normal speech comprehension and production. In contrast, they experience difficulties in recognizing instrumental melodies; they have problems hearing when someone sings out of tune or plays a "wrong" note (typically, a mistuned or out-of-key note); and the large majority sing out of tune. Amusics have difficulties recognizing hummed melodies from familiar songs, yet they can recognize the lyrics that accompany these melodies. In singing, they can recall the lyrics of familiar songs to which they can hardly produce a recognizable tune (e.g., Tremblay-Champoux et al. 2010).

Curiously, there is a paucity of research on this striking dissociation between music and speech. The only area of comparison studied so far concerns the intonation pattern of speech. In both French and English, intonation is used to convey a question or a statement. Amusics have little difficulty to distinguish these although they may show mild impairments when these pitch changes are subtle (Hutchins et al. 2010; Liu et al. 2010) or require memory (Patel et al. 2008b). Similarly, amusics may experience difficulties when comparing lexical tones taken from Mandarin or Thai (Tillmann et al. 2011). Speakers of a tonal language essentially show the same profile (e.g., Nan et al. 2010). Thus, amusics may show a deficit in processing pitch information in speech but this deficit is generally mild.

The clear-cut dissociation between music and speech seen in amusia provides a unique opportunity to address other fundamental questions related to the comparison of music and speech. For example, a current, hotly debated issue concerns the sharing (or overlap) of the processing involved in music and speech syntax. As mentioned above, a behavioral failure in the detection and discrimination of melodies by an out-of-key note is diagnostic of the presence of congenital amusia, presumably because the out-of-key note is tuned correctly but violates the tonal ("syntactic") relationships between notes in the given key of the melody. According to Patel's *shared syntactic integration resource hypothesis*" (SSIRH), discussed further below, amusics should exhibit similar

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.

difficulties with language syntax. Future research is needed to determine the analogous situation in language.

What counts as music or as nonmusical is not trivial. For example, rap music may be heard as speech, and highly dissonant music as noise. Conversely, some speech streams, such as the typical speech used by an auctioneer, may not be considered musical yet this form of chanting might be processed as music. Such ambiguous signals are not problematic for the peripheral auditory system, which does not need to decide which part of the auditory pattern is sent to music processors and which part to the language system. All information in the auditory input, including the text and the melody of an auction chant, would be sent to all music and language processors. The intervention of music- or language-specific components is determined by the aspect of the input for which the processing component is receptive. Thus, by studying the way amusics process different forms of music and speech, we may gain insight into which aspects are essential and specific to music.

Vocal control has also been studied in sensory and motor amusia (i.e., the loss or impairment of the ability to perceive or produce music or musical tones) (Ayotte et al. 2002; Loui et al. 2008, 2009). Diffusion tensor imaging has shown that in the right hemisphere of amusics, there is a thinner arcuate fasciculus: a fiber bundle connecting pars opercularis and the superior temporal areas, which is believed to provide auditory feedback control of speech. Amusics also have deficits in discriminating statements from questions (Liu et al. 2010) when there are small (4–5 semitone) pitch movements. The higher threshold for discriminating pitch movement in amusics may impair musical perception (that uses 1–2 semitone intervals) without usually affecting speech perception, which uses larger pitch movements (4–12 semitones). The areas affected are likely to include both the superior temporal and frontal areas. Evidence for domain specificity comes from patients with congenital amusia, who are impaired in short-term memory for music (pitch and timbre) but not for verbal material (Tillmann et al. 2009).

Beat Deafness

The most frequent form of congenital amusia affects the pitch dimension of music and spares, to some extent, rhythm. Recently, the reverse situation was observed in a young man in whom amusia is expressed by a marked difficulty to find and synchronize with the musical beat (Phillips-Silver et al. 2011). This case suggests a deficit of beat finding in the context of music. The subject is unable to period- and phase-lock his movement to the beat of most music pieces, and cannot detect most asynchronies of a model dancer (Phillips-Silver et al. 2011). The ability to identify or find beat has many practical uses beyond music and dance. For example, the act of rowing, marching in a group, or carrying a heavy object with others is made easier when beat is present.

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.

Nonmusical behaviors involving temporal coordination between individuals, such as conversational turn-taking or even simply adjusting one's gait to that of a companion, require sophisticated processes of temporal prediction and movement timing (see chapters by Levinson and Fogassi, both this volume). It is of interest to know whether such processes share mechanisms with beat finding. Does temporally coordinated behavior with another person outside of the musical domain rely on the same brain network involved in tracking and predicting beats in music? Future studies aim to test whether beat-deafness impacts speech rhythm, gait adjustment, or other nonmusical rhythmic tasks.

Animal studies may also be useful in elucidating the neural pathways involved in beat perception (Patel et al. 2009). Cockatoos, parrots, and elephants have been shown to be able to synchronize their movements to a musical beat. While there is also some evidence that monkeys display drumming behavior and that brain regions preferentially activated to drumming are also activated by vocalizations (Remedios et al. 2009), it has been shown that the same types of monkeys (rhesus macaques) cannot synchronize their taps to an auditory metronome (Zarco et al. 2009).

Brain Injury and Plasticity

One empirical approach that has been valuable both in the study of music and in the study of language is the evaluation of compensatory plasticity in the nervous system. Extreme cases of plasticity can be seen following stroke or a traumatic brain injury, or in developmental disorders of deafness and blindness. Both deafness and blindness lead to compensation of sensory loss by the remaining senses (cross-modal plasticity). Visual deprivation studies in animals and neuroimaging studies in blind humans have demonstrated massive activation of normally visual areas by auditory and somatosensory input (Rauschecker 1995). While changing sensory modality, formerly visual areas in the occipital and temporal lobe retain their functional specialization in the processing of space, motion, or objects, such as faces or houses (Renier et al. 2010). Restitution of functionality impaired after an insult is paralleled by micro- and macro-structural as well as representational (functional) changes in cerebral gray and white matter. These changes can be seen in the immediate perilesional cortex as well as in homologous regions in the unimpaired healthy hemisphere. The major mechanisms of this plasticity are regeneration and reorganization. Regeneration involves axonal and dendritic sprouting and formation of new synapses, most likely induced by the production and release of various growth factors and up-regulation of genetic regulators. Reorganization involves remapping of lesional area representations onto nonlesional cortex either in the perilesional region or in the contralesional hemisphere.

One of the most typical examples of lesional plasticity is the ability of the brain, through internal or external triggers, to reorganize language functions after an injury to the language-dominant hemisphere. The general consensus is

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 two routes to recovery. In patients with small lesions in the left hemisphere, there tends to be recruitment and reorganization of the left hemispheric perilesional cortex, with variable involvement of right hemispheric homologous regions during the recovery process. In patients with large left hemispheric lesions involving language-related regions of the frontotemporal lobes, the only path to recovery may be through recruitment of homologous language and speech-motor regions in the right hemisphere, recruitment which is most effective in young children. Activation of right hemispheric regions during speech/language fMRI tasks has been reported in patients with aphasia, irrespective of their lesion size. For patients with large lesions that cover language-relevant regions in the left hemisphere, therapies that specifically engage or stimulate the homologous right hemispheric regions have the potential to facilitate the language recovery process beyond the limitations of natural recovery. It is worth remembering that the plastic reorganization is age dependent and that some damage is irreversible.

Turning to the relation of music, melodic intonation therapy (MIT) is an intonation-based treatment method for severely nonfluent or dysfluent aphasic patients who do not have sufficient perilesional cortex available anymore for local functional remapping and reorganization. MIT has been developed in response to the observation that severely aphasic patients can often produce well-articulated, linguistically accurate utterances while singing, but not during speech. MIT uses a combination of melodic and sensorimotor rhythmic components to engage the auditory-motor circuitry in the unimpaired right hemisphere and trains sound-motor mappings and articulatory functions (Schlaug et al. 2010). In expressive aphasics, song may be used for therapeutic purposes to encourage the recovery of speech. MIT therapy has been used to help nonfluent aphasics recover speech, and it appears to work by recruiting neural plasticity in right hemisphere word articulation circuitry. Similar interventions for musical dysfunctions or the use of language structures and language tools to overcome musical dysfunctions have not been developed, but this could be an interesting line of research to pursue.

Role of Temporal Frontal Neuroanatomical Connections in Speech and Music Production

The left "perisylvian" cortex (consisting of superior temporal, inferior parietal, and inferior frontal regions) is seen as crucial for language perception and production, with various fiber pathways (see below) connecting the left superior temporal cortex ("Wernicke's area") and the left inferior frontal cortex ("Broca's area"). If someone suffers a large left hemispheric lesion, leading to aphasia, then the variability of the size of the right hemispheric language tracts might actually contribute to natural recovery of language function.

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.

Rilling et al. (2008; see also Figure 9.6 in Hagoort and Poeppel, this volume) have presented a comparative analysis of arcuate fasciculus (AF) across three species (macaque, chimpanzee and human). In all three cases there is significant connectivity along dorsal projections. However, the extensive ventral stream projections observed in the human brain is not observed in either the chimpanzee or macaque brain.

Three different tracts connect the temporal lobe with the frontal lobe: the AF, the uncinate fasciculus, and the extreme capsule. Most is known about the AF, which connects the STG and middle temporal gyri (MTG) with the posterior inferior frontal lobe, arching around the posterior Sylvian fissure. Recent studies have suggested that the AF may be primarily involved in the mapping of sounds to articulation (in singing and spoken language) and/or to audiomotor interactions in learning and performance of instrumental music. (Earlier, we suggested it provides auditory feedback control of speech.) Some believe that the AF is direct and that there are fibers between the STG/MTG and the inferior frontal gyrus (IFG) (Loui et al. 2008, 2009; Schlaug et al. 2010), whereas Frey et al. (2008) argue that the AF is an indirect tract, as in most nonhuman primates (Petrides and Pandya 2009).

The temporal component connects to the parietal lobe, and then the superior longitudinal fasciculus connects the parietal lobe to the IFG. The AF has connections with inferior primary somatosensory, inferior primary motor, and adjacent premotor cortex. In humans, the AF is usually larger in the left than in the right hemisphere, although the right hemisphere does have a complete tract which might allow the right hemisphere to support vocal output even if the left hemisphere is lesioned. In chimpanzees, arcuate terminations are considerably more restricted than they are in humans, being focused on the STG posteriorly and on the ventral aspects of premotor cortex (BA 6) and pars opercularis (BA 44) anteriorly. In macaques, the arcuate is believed to project most strongly to dorsal prefrontal cortex rather than to Broca's area homologue.

Schlaug and colleagues have shown that the AF of musicians is larger in volume than in nonmusicians (Halwani et al. 2011); it also differs in microstructure (fractional isotropy) from nonmusicians. Moreover, in singers, the microstructural properties in the left dorsal branch of the AF are inversely correlated with the number of years of vocal training. These results suggest that musical training leads to long-term plasticity in the white matter tracts connecting auditory-motor and vocal-motor areas in the brain. To complicate matters, there may be a developmental story in which myelination and maturation of these fiber bundles in the AF influences language development (Brauer et al. 2011a).

The uncinate fasciculus is a hook-shaped fiber bundle that links the anterior portion of the temporal lobe with the orbital and inferior frontal gyri. The extreme capsule is a fiber bundle that links the temporal with more anterior portions of the IFG (Brodmann 45) and inferior prefrontal regions. Both the

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.

uncinate fasciculus and the extreme capsule are thought to be more involved in the mapping of sounds to meaning. Both fiber tracts can carry information along the ventral "what" pathway from the anterior STG to IFG (Marchina et al. 2011). We speculate that these ventral pathways are likely to be important for the processing of speech meaning whereas dorsal pathways are likely to be important in speech and musical production.

Ventral pathways are perceptual; they allow auditory object identification and association with behavioral "meaning." Dorsal pathways associate sounds with actions. In Hickok and Poeppel's model, area SPT is involved in the transformation of perceived and spoken words (Hickok and Poeppel 2004); Rauschecker and Scott (2009) emphasize the importance of the reverse transformation. Unresolved is how the architecture of the multistream models that have provided useful heuristics in speech and language research extends to music. Clearly these models are integrated and "deployed" during song, since speech/language are half the battle. However, since these models are also motor control and sensorimotor transformation models, it stands to reason that they also play a central role in performance of instrumental music, and crucially in the predictive aspects of processing. (For a complementary view of the dorsal and ventral pathways, in this case in the visual control of action, see Arbib et al., this volume. The view there is that the dorsal pathway is implicated in the parameterization of action whereas the ventral pathway can invoke object identification to support prefrontal planning of action.)

Speech and Song Production

Speech production mechanisms are intimately tied to song production (as discussed further by Janata and Parsons, this volume). Speech and language production involves a multistage process: first you must select an appropriate message, then each lexical item (a lemma) to express the desired concept, and then access the sound structure. Of course, additional stages are also necessary for the construction of hierarchically organized sentences or intonation contours. Brain activation (reviewed in detail by Indefrey 2011) includes sensory-related systems in the posterior superior temporal lobe of the left hemisphere; the interface between perceptual and motor systems is supported by a sensorimotor circuit for vocal tract actions (not dedicated to speech) that is very similar to sensorimotor circuits found in primate parietal lobe (Rauschecker and Scott 2009). The posterior-most part of the left planum temporale (SPT) has been suggested to be an interface site for the integration of sensory and vocal tract-related motor representations of complex sound sequences, such as speech and music (Hickok and Poeppel 2004, 2007; Buchsbaum et al. 2011). As such, SPT is part of a dorsal-processing stream for sensorimotor control and integration, where general sensorimotor transformations take place for eye and limb movements in the service of internal models of behavior and optimal state control

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.

(Rauschecker and Scott 2009). The data cited above on vocalization and singing suggest that song without words builds on the same circuitry (Zarate et al. 2010; Zarate and Zatorre 2008).

A comparison of speech and singing (Özdemir et al. 2006) shows shared activation of many areas, including the inferior pre- and postcentral gyrus, STG, and STS bilaterally. This indicates the presence of a large shared network for motor preparation and execution as well as sensory feedback/control for vocal production. Hence, these results suggest a bi-hemispheric network for vocal production regardless of whether words or phrases were intoned or spoken. However, singing more than humming ("intoned speaking") showed additional right-lateralized activation of the STG, inferior central operculum, and IFG. This may explain the clinical observation that patients with nonfluent aphasia due to left hemisphere lesions are able to sing the text of a song while they are unable to speak the same words. The discussion on melodic intonation therapy above provides an important connection point here.

Potential Right Hemisphere Biases: Evidence from Neuropsychology and Neuroimaging

Based on neurological cases and neuroimaging research, evidence suggests that musical pitch perception, or at least dynamic pitch, has a right hemisphere bias in the auditory cortex. There is now evidence (Foster and Zatorre 2010) that cortical thickness in right Heschl's sulcus and bilateral anterior intraparietal sulcus can predict the ability to perform relative pitch judgments. The intraparietal sulcus is known to play a role in other transformations, since it is activated during visuospatial rotation (Gogos et al. 2010) and mental melody rotation (Zatorre et al. 2010). However, there is not universal agreement on the role of the right hemisphere in pitch; for example, there are differences in hemispheric contributions in absolute pitch and nonabsolute pitch musicians (Brancucci et al. 2009).

Left temporal areas have been shown to be important for fine intensity discrimination and fine pitch discrimination (Reiterer et al. 2005); right temporal areas are more important for other highly differential acoustic stimuli (i.e., holistic feature processing). However, in a more careful parametric study, Hyde et al. (2008) showed that the right hemisphere has higher pitch resolution than the left hemisphere. Left auditory cortex also showed greater activation during active stream segregation (Deike et al. 2010). In an interesting study that employed two discrimination tasks (tone contour vs. duration) with identical stimuli in each task condition (Brechmann and Scheich 2005), the right hemisphere auditory cortex was more strongly activated for the contour task, whereas the left hemisphere auditory cortex was more strongly activated for the duration task. It is important to emphasize, however, that the auditory cortices were bilaterally activated in both task conditions. These results indicate that there is no

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.

simplistic right or left hemisphere specialization in general auditory analysis and that, even for the same acoustic stimuli, lateralization may vary with task condition and demands.

Laterality in processing of vocalizations appears to have emerged early in evolution in primates and avian species. Neuronal responses to vocalizations in primates have been described in a network that includes STS, STG, and temporal pole, cingulate, and inferior frontal cortex. Laterality has been described in monkeys based upon imaging and lesion studies (Heffner and Heffner 1984; Harrington et al. 2001; Poremba et al. 2004; Joly et al. 2012).

Klein and Zatorre (2011) investigated categorical perception, a phenomenon that has been demonstrated to occur broadly across the auditory modality, including in the perception of speech (e.g., phonemes) and music (e.g., chords) stimuli. Several functional imaging studies have linked categorical perception of speech with activity in multiple regions of the left STS: language processing is generally left hemisphere dominant whereas, conversely, fine-grained spectral processing shows a right hemisphere bias. Klein and Zatorre found that greater right STS activity was linked to categorical processing for chords. The results suggest that the left and right STS are functionally specialized and that the right STS may take on a key role in categorical perception of spectrally complex sounds, and thus may be preferentially involved in musical processing. It is worth noting, however, that not all phonemes are categorically perceived; for instance, vowels and lexical tones of tone languages do not have categorical perception, although they are stable sound categories (Patel 2008). Conversely, there is evidence for categorical perception of tone intervals in musicians (Burns and Ward 1978).

Domain-Specific Processes: Neural Substrates for Speech

Speech Perception

Two stages can be identified in the perception of speech: phonological information (i.e., speech sounds) must be recovered and lexical-semantic information must be accessed. The recognition of speech sounds is carried out bilaterally in the superior temporal lobe (with a left hemisphere bias); the STS is bilaterally (and increasingly anteriorly) involved in phonological-level aspects (phonemes, words, and short phrases) of this process (DeWitt and Rauschecker 2012). The frontal premotor system is not involved in the perception of speech sounds per se (i.e., decoding of sounds and speech recognition in naturalistic conditions), but is important for their categorization in laboratory tasks. Currently it is unclear where conceptual access mechanisms are located in the brain, although the lateral and inferior temporal lobes (middle and inferior temporal gyri) most likely play a role.

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.

Differences between Neural Substrates for Language and Music

Within-area differences have been found between activation for speech and music. Multivariate pattern classification analyses (Rogalsky et al. 2011b) indicate that even within the regions of blood oxygenation, level-dependent (BOLD) response overlap, speech and music elicit distinguishable patterns of activation. This raises the possibility that there are overlapping networks or even distinct and separate neural networks for speech and music that coexist in the same cortical areas. Such a view is supported by a recent fMRI study which defined language regions functionally in each subject individually and then examined the response of these regions to nonlinguistic functions, including music; little or no overlap was found (Fedorenko et al. 2011). However, as Patel (2008) observes, other studies show that musical training influences the cortical processing of language (Moreno et al. 2009) and supports the idea that there are shared networks, as seems obvious at least for "early" auditory regions.

Activation for sentences and melodies were found bilaterally in the auditory cortices on the superior temporal lobe. Another set of regions involved in processing hierarchical aspects of sentence perception were identified by contrasting sentences with scrambled sentences, revealing a bilateral temporal lobe network. Sentence perception elicited more ventrolateral activation, whereas melody perception elicited a more dorsomedial pattern, extending into the parietal lobe (Rogalsky et al. 2011b).

Patel (this volume) offers the "dual systems" SSIRH model to explain the domain-specific representations in long-term memory (i.e., stored knowledge of words and their syntactic features and stored knowledge of chords and their harmonic features) and shared neural resources that act on these representation networks (see also Patel 2003, 2011). However, although there is considerable support for the SSIRH model, there is some controversy over the degree of shared neural resources for syntactic processing in music and language. For example, Maidhof and Koelsch (2011) examined the effects of auditory selective attention on the processing of syntactic information in music and speech using event-related potentials. They suggest that their findings indicate that the neural mechanisms underlying the processing of syntactic structure of music and speech operate partially automatically and, in the case of music, are influenced by different attentional conditions. These findings, however, provide no clear support for an interaction of neural resources for syntactic processing already at these early stages. On the other hand, there is also evidence for shared mechanisms. When an acoustic (linguistic or musical) event occurs that violates the expectations of the predictive model, the brain responds with a powerful mismatch response. This can take the form of mismatch negativity for oddballs or violations of acoustic patterns, and may lead to bi-hemispheric changes

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.

in the evoked brain potentials: early left anterior negativity (ELAN) for the presentation of an unexpected syntactic word category and early right anterior negativity (ERAN)) after the presentation of a harmonically unexpected chord at the end of a sequence. The SSIRH model (Patel 2003) and several recent studies suggest that linguistic and musical syntax may indeed be co-localized and overlapping (Sammler et al. 2012).

Outlook: Challenges and Mysteries

Dance and Music

Janata and Parsons (this volume) provide a discussion of the neural mechanisms involved in music, song, and dance. To focus our efforts at the Forum, we limited our discussion primarily to a consideration of spoken rather than signed language. Still, we emphasize the importance of gesture and movement in language, both as a vibrant accompaniment to spoken language and as a signal in conversational turn-taking, in musical performance as well as in dance. Future research will need to address the dimension of body movement and integrate it with our understanding of music. In particular, it will be very important to learn how kinesthetic, proprioceptive, and visual cues are integrated with the motor and auditory systems.

Poetry and Song: Bringing Music and Language Together

"Language" and "music" actually form two poles of a continuum that includes song-like or musical speech, tonal languages, poetry, rap music, and highly syntactically structured music. Lewis (this volume) describes a fusion of language and music in the BaYaka Pygmy hunter-gatherers in the Congo, and Levinson (this volume) observes that "song in a sense is just language in a special, marked suprasegmental register or style or genre" and that "music may be an ethnocentric category" (Nettl 2000).

Parallel to the controversies over the neural representation for language and music mentioned above, there continues to be vigorous debate about the relationship between the processing of tunes and lyrics in song, and about the neural structures involved. While there is good neuroimaging and neuropsychological evidence for separate processing of lyrics and melody in song, there is also compelling evidence for integrated processing of words and music in a unified neural representation. While brain activation patterns evoked by the perception and production of song show overlap with the spoken word activation network in many studies (Janata and Parsons, this volume), other studies emphasize differences. Patel (this volume) has suggested that there is a song sound map in the right hemisphere and a speech sound map in the left hemisphere. Experimental support for such hemispheric specialization at a

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.

production level is provided by a study which found that the right IFG, right premotor cortex, and right anterior insula were active in singing only, suggesting that song production engages some right hemisphere structures not activated in normal speech (Saito et al. 2006). Right hemisphere dominance for singing has also been shown by TMS studies (Stewart et al. 2001). However, activation studies of song perception by Schön et al. (2010) and adaptation studies of Sammler et al. (2010) argue against domain specificity and show broad, bilateral activation of auditory areas in superior temporal lobe and STS for lyrical songs. In more detail, the latter results also show greater integration of lyrics in tunes in the left middle left STS, suggesting lyrics and tunes are strongly integrated at a prelexical, phonemic level. The more independent processing of lyrics in the left anterior STS may arise from analysis of meaning in the lyrics. An explanation for divergent and disparate reports in the literature may be that there are variable degrees of integration/dissociation of lyrics and melody at different stages of song perception, production, and memory (Sammler et al. 2010). Depending on the specific cognitive demands of an experimental task, text and melodies may be more or less strongly associated but not fully integrated, and the extent of integration may also vary with the degree of familiarity of the song to the listener, and the listener's attentional focus. Additional variation can also occur in other ways within the same song; for example, vowels are more tightly bound with pitch information than consonants in song perception (Kolinsky et al. 2009). There may be more variation and independent processing at the perceptual rather than the production level since lyrical and melodic features of song must be integrated in the output stage as a vocal code for singing.

Additional Problems and Challenges for Future Research

Our search for the neural and computational "primitives" underlying music and language, "domain-specific" and "domain-general" representations and computations, and our summary of current neurobiological insights into the relations between language and music have revealed a tremendous, recent surge of research and interest in this interdisciplinary field, and yielded an extraordinary treasure trove of fascinating advances, many achieved with dazzling new neuroscientific techniques. For example, we have described great advances in understanding brain development and plasticity during acquisition of language and music, insights into the neural substrates of emotional responses to music (Salimpoor et al. 2011), the relation between music and language perception and production in the perception–action–prediction cycle, the evidence for separable modular components for speech and music processing both at lower auditory levels and a higher cognitive level. Although there is compelling neuropsychological data for a neat dissociation between the neural substrates for music and language, the neuroimaging data tell a more complex story. While

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.

many neuroscientists think of music and language as distinct modular systems, another viewpoint is that they are different ends of the continuum of "musilanguage" that also includes song and poetry (Brown 2000), with music emphasizing sound as emotional meaning whereas language emphasizes sound as referential meaning. Given the range of perspectives in this field, and the fundamental questions that still remain unanswered, it is clear that are still many "gaps" in our knowledge. Thus, in an effort to spur future research, we conclude by listing areas that we feel require further study:

1. How are the memory systems for language and music both independent and interwoven? How are the lyrics and melody of familiar songs separately and conjointly stored?
2. What are the parallel and overlapping substrates for language and music acquisition during childhood development? Do structural and functional brain changes occur during the learning of speech and music?
3. What are the shared versus distinct speech and song production mechanisms?
4. What causes lateralization? Is there an overall right hemisphere lateralization for music and left hemisphere lateralization for speech?
5. What are the neural representations and multisensory mechanisms shared by dance, music, and language? Is there a common neural basis underlying the ability of dance, music, and language to evoke emotions?
6. Precisely what contributions do brain oscillations make to auditory processing in language and music? How best can these influences be explored, evaluated, and critically tested?
7. How have speech and music evolved through the prism of animal models of communication and rhythm perception?
8. What is the nature of the interaction between external acoustic inputs and anticipatory and predictive internal feedforward systems in language and music during conversation and improvisation?

Acknowledgment

We gratefully acknowledge the contributions of Petr Janata to our discussions.

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.