Linguistics and Philosophy
https://doi.org/10.1007/s10988-022-09365-y

ORIGINAL RESEARCH

Musical grouping as prosodic implementation

Jonah Katz[1]

Accepted: 12 July 2022

Abstract
This paper reviews evidence concerning the nature of grouping in music and language and their interactions with other linguistic and musical systems. I present brief typological surveys of the relationship between constituency and acoustic parameters in language and music, drawing from a wide variety of languages and musical genres. The two domains both involve correspondence between auditory discontinuities and group boundaries, reflecting the Gestalt principles of proximity and similarity, as well as a nested, hierarchical organization of constituents. Typically, computational-level theories of musical grouping take the form of a function from acoustic properties through grouping representations to syntactic or interpretive constituents. Linguistic theories tend to be cast as functions in the opposite direction. This study argues that the difference in orientation is not grounded in principled differences in information flow between the two domains, and that reconceptualizing one or both theories allows for gains in analytical understanding. There are also obvious differences between musical and linguistic grouping. Grappling with those differences requires one to think in detail about modularity, information flow, levels of description, and the functional nature of cognitive domains.

Keywords Prosody · Grouping · Gestalt · Music · Language · Modularity

1 Introduction

This paper reviews similarities and differences between linguistic prosodic phrasing (referred to here as *grouping*) and musical grouping, including their interactions with other components of musical and linguistic structure. Such comparisons are of interest in part because they bear on the question of *modularity*: to what extent do various cognitive processes apply to different sensory inputs, drive different kinds of behavior, and make use of different representations? There is no shortage of review literature on the music-

Many thanks to Salvador Mascarenhas (the action editor for this paper), Michael Wagner, and two anonymous reviewers for helpful discussion of some of this material. This paper is part of the Special Issue "Super Linguistics", edited by Pritty Patel-Grosz, Emar Maier, and Philippe Schlenker.

Jonah Katz
jokatz@mail.wvu.edu

[1] Department of World Languages, Literatures and Linguistics, West Virginia University, Morgantown, WV, USA

Published online: 26 October 2022

language comparison, from a variety of different methodological and theoretical perspectives, and the current paper does not comprehensively summarize all available research. Instead, I explore one particular aspect of music–language similarity, the structure and function of grouping, and from this exercise draw some conclusions about the domain-generality of grouping principles. One crucial point is that *music* and *language* each constitute a complex and heterogeneous set of representations, processes, and behaviors. Each of these components may rely on information from other components internal or external to their cognitive domains. And each component may be described at different levels. Progress in understanding the relationship between music and language requires clarifying the details of implicit computational principles underlying specific processes, representations, and behaviors in the two domains.

1.1 Theoretical and methodological preliminaries

One of the central questions in the history of cognitive science is the extent to which language makes use of information in a manner distinct from other cognitive domains. The term *modularity* has several other implications (e.g. brain localization, automaticity), but I use it here to single out this particular question. While it may seem straightforward to examine existing theories of language and music and compare their properties, there are a number of pitfalls inherent to the enterprise. Because neither language nor music can be rigidly defined by association with a specific set of behaviors or computations, the only coherent way to think about either domain is as a collection of heterogeneous cognitive resources. If we find that many of these resources are common to music and language, it may imply that they are domain-general or it may imply that *music* and *language* are words used to describe overlapping sets of cognitive domains.

When comparative study suggests that music and language share some property, it raises several questions. One is what kind of a property it is (e.g. structural, behavioral, neural). A second question is why the property is the way it is, instead of being some other property. A third question is why the two domains might share the property. The variety of possible answers to such questions implies a variety of different forms of domain-generality that could be of interest to cognitive scientists. Two domains might share some resource for 'low-level' reasons related to some more basic domain, for reasons related to the internal structure of the domains themselves, or, possibly, by coincidence. Any such parallels may be the consequence of homology, where two domains share properties because they are deeply related at the level of human biology; or analogy, where the property arises independently more than once because it is a 'good solution' to some problem.

Marr (1982) distinguishes three levels of description for any cognitive process. The *computational* level describes in abstract symbolic terms what is being processed, what the process itself is like, and why the process looks the way it does instead of some other way. The *algorithmic* level describes how that process is applied to input representations in real time. And the *implementational* level describes the machinery that performs the algorithm, generally part of the brain.

Marr emphasizes the fact that descriptions at the three levels are necessarily related to one another in intricate and complex ways, but the relationship is not deterministic: some analytical choices at one level are independent of choices at other levels. This is important because it also has implications for the study of shared resources between cognitive domains. Music and language may share computational properties, but implement them

with different algorithms in different areas of the brain, or with the same algorithm in different areas of the brain. Similarly, they may share some processing or learning algorithm but apply it to completely different representations and thus generate systems with different computational properties. In this paper, I focus on the computational level, partly out of the conviction that if we are to understand parallels in processing or neural substrates, we first must understand *what* is being processed. Heffner and Slevc (2015) offer an overview more oriented towards processing and neuroimaging.[1]

1.2 Grouping and Gestalt principles

Gestalt grouping principles are the main component of Wertheimer's (1938) approach to perception. Wertheimer was mainly concerned with vision in his work, and the grouping principles were attempts to describe which stimulus features result in some parts of a visual array being perceived as forming a unit (*group*) to the exclusion of others. A linguist might say that the Gestalt principles are a theory of visual constituency and how it is perceived. This paper is mainly concerned with two Gestalt principles referred to as *similarity* and *proximity*. These are illustrated in Fig. 1. Most illustrations of visual Gestalt principles use simple geometric shapes such as circles or triangles; I use more complex shapes for reasons that will be clarified momentarily.

In Fig. 1a, a series of identical zig-zag shapes is evenly spaced on the page. The claim here is that no particular grouping inheres across shapes: a viewer may divide them into groups explicitly or implicitly, perhaps in a way that creates a small number of groups of roughly equal cardinality, but there is no particular cue to reinforce any specific grouping. In (1b), on the other hand, the first 3 shapes are closer together than the third and the fourth, and the last 4 shapes are also closer together. Here, the claim is that viewers will perceive two groups consisting of shapes 1–3 and shapes 4–7. This is the proximity principle: objects that are closer together are more likely to be perceived as a group. In (1c), all of the shapes are again equally spaced, but now shapes 4–7 are smaller than 1–3. Another way to state this is that the parameter 'size' changes between shapes 3 and 4 in a way that it does not change between any other adjacent pair. This is the similarity principle: objects that are more similar are more likely to be perceived as a group. Here I have manipulated similarity in *size*; Gestalt theory claims that other visual parameters, such as color or shape, also trigger similarity-based grouping inferences.

While much of the original work in Gestalt psychology concerned vision, the approach is meant as a general framework for perception, and it has been extended to other domains and modalities. Wagemans et al. (2012) offer a detailed historical and theoretical overview of visual grouping, Denham and Winkler (2014) review Gestalt theories of auditory perception, and Gallace and Spence (2011) review evidence for Gestalt principles in multiple modalities including tactile perception. The current paper focuses on auditory grouping. To understand the auditory analogs of the proximity and similarity principles in Fig. 1, imagine that the visual shapes therein are representations of air pressure (vertical dimension) over time (horizontal); that is, sound waves, as they might be viewed in acoustic software. In this case, (1a) would represent a sequence of identical sounds

[1] There are several related topics that I don't have space to discuss here. I briefly touch on meter and syntax, but only to the extent that they interact with grouping. Each of those components could be the subject of a book-length review on its own. For interesting overviews, see Fitch (2015) on meter and Rohrmeier et al. (2015) on syntax. I also mention musical textsetting in a few examples, but do not address it in any detail; Dell (2015) gives a highly relevant overview.

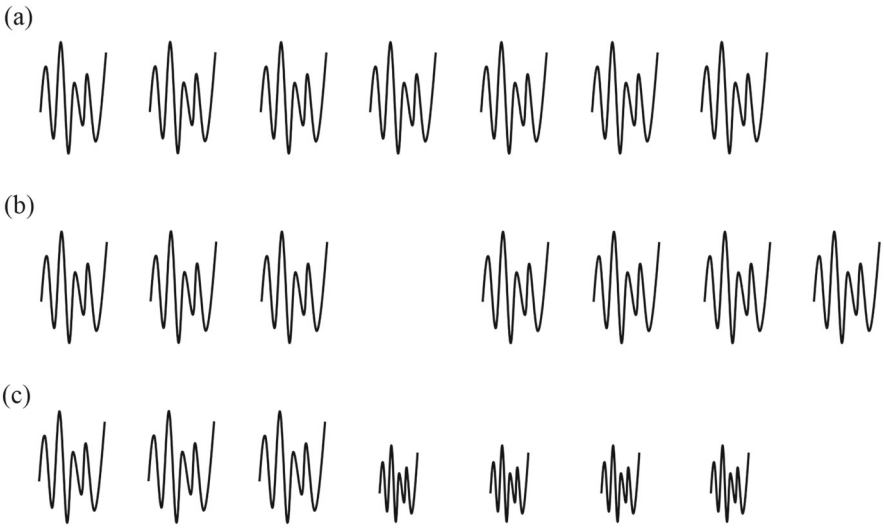

Fig. 1 Proximity and similarity in visual arrays: **a** ungrouped stimuli; **b** proximity-based grouping; **c** similarity-based grouping on the size dimension

recurring at perfectly isochronous intervals. (1b) would represent three sounds separated by short pauses, followed by a long pause, then four more sounds separated by short pauses. (1c) would represent three relatively high-amplitude sounds (correlated with the perception of loudness) followed by four lower-amplitude sounds. The claim of Gestalt theory is that the perceived grouping of these auditory stimuli would be the same as their visual counterparts: the proximity principle for audition deals with proximity in *time*, and the similarity principle can apply to any auditory parameter, such as loudness, pitch, or timbre (correlated with spectral characteristics). As Wagemans et al. (2012) point out, the proximity principle is often treated separately from the similarity principle, but it is in fact a special case of similarity in spatial position (for vision) or temporal onset (for audition).

There is a fairly clear intuition about why auditory Gestalt principles work the way they do. In general, grouping inferences based on proximity and similarity are relatively likely to accurately reflect the sources of sounds in the environment (e.g. Deutsch, 1999; Schlenker, 2017). For instance, two sequences of sounds separated by a pause are more likely to come from two different sources than two sequences not separated by a pause. The same is ostensibly true for sequences separated by a discontinuity in pitch, intensity, etc. The generality of this reasoning is one reason why Gestalt principles seem to apply across such a wide variety of domains. Signed languages implement prosodic groups visually using the proximity rule (mainly for manual signs) and possibly similarity rules (for non-manual signs; see Fenlon and Brentari, 2021 for an overview). Charnavel (2019, 2022) and Patel-Grosz et al. (2018) argue that the structure of dance also makes use of Gestalt visual grouping. And Spelke (1994) explains why certain principles of spatial cognition and object recognition, some of which are related to proximity and similarity, are likely to give rise to ecologically valid inferences about objects in the world. In this paper, I focus on Gestalt principles in music and spoken language.

Note that the focus here is on grouping of sounds across time, into a series of constituents. I refer to this as *rhythmic* grouping, to distinguish it from a different use of the term in audition. Some auditory researchers use the term *grouping* to refer to the process of

separating a complex auditory signal with multiple overlapping sounds into component 'voices' or 'streams'. While this activity is clearly relevant to both speech perception (e.g. Bregman, 1990) and music (e.g. Huron, 1991), and is frequently discussed in terms of the same Gestalt principles used here, I make no strong claims about this type of stream segregation. Instead, I use the term 'grouping' to refer to the rhythmic variety. The question of whether the two types of grouping rely on the same Gestalt principles is an interesting and complex one, which will not be addressed here.

1.3 A brief outline

In Sects. 2 and 3, I outline some characteristics of grouping across languages (generally referred to as *prosodic phrasing*) and genres of music. These descriptions suggest that the two systems are not only similar in their internal organization, but that they also display *substantive* similarities in the types of acoustic discontinuity and change that inhere to groups in the two domains, related to Gestalt principles. I use the terms *discontinuity* and *disruption* interchangeably, to describe situations where some acoustic parameter changes from one event (note, syllable, segment, etc.) to the next, and where that same parameter displays less of a change before and after the events in question, as in Fig. 1b-c. The similarities in substantive and formal properties of groups in language and music raise the question of whether musical grouping and its relationship to harmonic 'syntax' (here, tonal and harmonic structure) can profitably be studied using tools from the syntax-prosody interface in linguistics.

One obstacle to such an undertaking is the fact that theories of musical grouping tend to take the form of a function from acoustics to groups, and a function from groups to constituents involved in memory, chunking, or harmonic computations. This is a 'reversal' of standard linguistic theories, which tend to take the form of a function from syntax and pragmatics to prosodic groups, and a function from prosodic groups to acoustic properties. In Sect. 4, I argue that the directionality in the musical functions is not motivated by considerations of information flow, and that it is plausible to think of Gestalt similarity and change principles as generating acoustic properties on the basis of grouping, much as they do in language.

The result of applying this linguistic perspective to musical 'syntax' and 'prosody' brings to the fore some *differences* between the two modalities: in particular, the representations to which prosodic implementation is applied must be rather different in the two domains. I argue that these differences follow from linguistic duality of patterning and, ultimately, from the nature of the linguistic lexicon. In language, syntactic primitives (bundles of features) bearing a law-governed relationship to semantics are spelled out in terms of arbitrary sound features with no law-governed relationship to syntactic features or meaning; this is the nature of the linguistic lexicon. In music, on the other hand, syntactic primitives are 'spelled out' with far fewer constraints on acoustic parameters, because there is no lexicon to constrain, for instance, metrical or intensity properties. The approach offers a new perspective on why musical groups are the way they are instead of some other way, as well as the similarities and differences between language and music.

2 Grouping in language

Speech and music both involve sequences of auditory events in time. Some of those events pattern together to the exclusion of others with regard to perceived coherence, acoustic patterns, or production regularities. I follow Lerdahl and Jackendoff (1983) in referring to this topic as *grouping*. Linguists generally call this structure *prosodic phrasing* (a term which also carries additional implications beyond perceptual coherence), but I use *grouping* here for consistency. In this section, I briefly review the overarching theory and some specific acoustic reflexes of grouping in language.

Linguistic theories of grouping tend to be concerned with relating two sets of empirical results, one from sound patterns and one from syntactic patterns. The first set of results involves context-dependent phonetic or phonological realizations of particular classes of sounds. Such patterns often reveal that some sounds pattern together to the exclusion of others, that is, phonetics and phonology operate on *constituents*. The second set of results involves the same observation, *mutatis mutandis*, for syntactic patterns: syntactic generalizations are also necessarily stated over constituents. The goal of linguistic theories of grouping is to explain the ways in which these two sets of facts are related: sound constituents can only be described with reference to (among other things) syntactic structure (Pierrehumbert, 1980; Selkirk, 1972). The dominant framework for unifying these results is the *prosodic hierarchy* (Nespor and Vogel, 1986; Selkirk, 1984). This approach characterizes the sound reflexes of grouping as resulting from: (1) a function that maps from syntactic structure to groups; and (2) a function that maps from mental representations of grouped sounds to the pronunciation of those sounds. As such, prosodic phonology requires a theory of groups, a theory of the function from syntactic structure to groups, and a theory of the function from groups to sounds.

2.1 Broad characteristics of linguistic grouping representations

A variety of factors go into determining the grouping structure of an utterance: syntax, pragmatics (including *information structure*), phonological quantity, speech rate, and affect, for instance, all play a role (Shattuck-Hufnagel and Turk, 1996). We focus here on the role of syntax, because it is relatively uncontroversial and well studied. The general 'shape' of grouping structures in language involves constituents nested hierarchically inside larger constituents (Hayes, 1989; Nespor and Vogel, 1986; Selkirk, 1984). For example, some token of the English utterance *the child in front discovered an object* might be grouped as in Fig. 2. Grouping is shown in two formally equivalent notations here: the bracketing generally used in linguistics (though without any notation of prominence) and the slur notation introduced by Lerdahl and Jackendoff (1983) in their exposition of musical grouping. This is to reinforce the point that, though theories in these two domains use different visual conventions to represent grouping, the information they are representing is exactly the same.

There are several important caveats here. First, the mapping from syntax to grouping is not deterministic, so a range of slightly different grouping structures is possible for this utterance. Second, theories differ as to how many distinct levels of grouping are present in such an utterance, and in details of where the boundaries between groups fall. Third, representations of linguistic grouping generally also display headedness, which I omit here but discuss in Sect. 4: the head of a prosodic group is the most structurally prominent unit in that group, as inferred from properties such as phonological stress and accent. Finally,

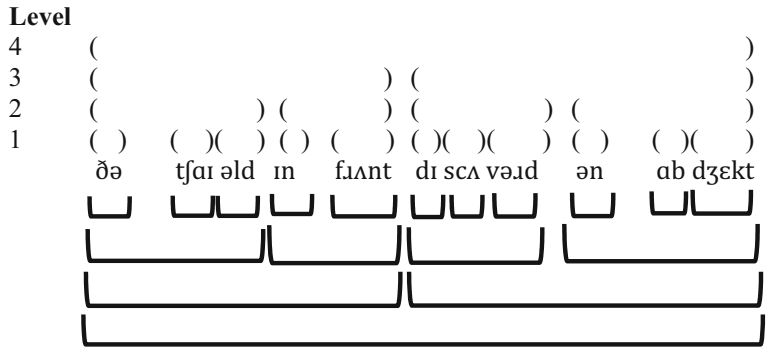

Fig. 2 A possible prosodic (grouping) realization of the English utterance *the child in front discovered an object*. The same grouping structure is shown in bracket notation (above the text) and Lerdahl and Jackendoff's (1983) grouping notation (below the text).

most linguistic theories posit labels for the different levels shown in Fig. 2. That is, groups at the same level are posited to be the same *type* of group, and sound patterns are generally described in terms of those types. Typical labels here might be: (1) *syllable*, (2) *prosodic word*, (3) *phonological phrase*, and (4) *utterance*, though the number and nature of these labels differ between theories.

In this example, I leave out the traditional labels. This reflects an emerging consensus that prosodic groups are not as neatly layered as originally believed, instead closely mirroring syntactic structure (Féry and Truckenbrodt, 2005; Ishihara, 2003; Ladd, 1986). The resultant theories either weaken (Féry, 2010; Selkirk, 2011) or eliminate (Wagner, 2005) the distinctions between various levels of the prosodic hierarchy.

Despite differences between theories, most researchers agree on several broad formal properties of grouping structure in language. Phonetic material is exhaustively partitioned into groups. If a group *X* contains some of the elements in another group *Y*, then either *X* or *Y* contains all of the elements in the other group (no partial overlap). Every utterance coincides with a single group that contains all other groups. And there is a tendency for groups to contain exactly two smaller groups. These general representational properties of linguistic grouping are inferred on the basis of sound properties from various languages. Next, we turn our attention to the nature of those properties.

2.2 The phonetics of linguistic grouping

A wide variety of phonetic and/or phonological generalizations have been described with reference to grouping structure. In this section, I focus on three phonetic dimensions: duration, pitch, and consonant manner. Details of all three dimensions depend on the context in which a sound is uttered, and stating those contexts requires reference to grouping structure.

2.2.1 Duration

The duration of a sound depends on many factors, including its inherent features, speech rate, the sounds that occur around it, and stress. Even in the face of this pervasive variation, however, it is possible to identify strong tendencies in how grouping affects duration.

Many languages lengthen the final element within a group. This *final lengthening* occurs at various levels of group boundary, and affects a variety of sounds located near such boundaries. Wightman et al. (1992) find successively longer vowels at the ends of English words, small prosodic groups, and larger groups. Lengthening at the ends of English words compared to word-medial vowels, on the other hand, is harder to detect and may be limited to vowels near pitch accents (Turk and Shattuck-Hufnagel, 2000). Gordon and Munro (2007) provide evidence for final lengthening of Chickasaw vowels at the utterance, phrase, and (perhaps) word levels. They also review other languages where final lengthening is attested at one or more levels, including Arabic (De Jong and Zawaydeh, 1999), Finnish (Oller, 1979), Greenlandic Eskimo (Nagano-Madsen, 1992), Mandarin (Duanmu, 1996), Yoruba (Nagano-Madsen, 1992), and Creek (Johnson and Martin, 2001).

Many languages also lengthen the *initial* element within groups. This mainly affects consonants, and is true of both articulatory and acoustic measurements. Consonants show initial acoustic and/or articulatory lengthening at one or more levels of grouping in Korean (Jun, 1993), English (Fougeron and Keating, 1997), French (Keating et al., 2003), Gurindji (Ennever et al., 2017), Taiwanese (Keating et al., 2003), Campidanese Sardinian (Katz and Pitzanti, 2019), Spanish (Kingston, 2008), and Japanese (Onaka et al., 2003). These initial duration differences are generally accompanied by a suite of articulatory effects referred to as *initial strengthening*; I leave this aside until the discussion of consonant manner in Sect. 2.2.3.

In general, then, sounds at the initial and final edges of groups tend to be longer than their counterparts internal to groups. This is not true for every speaker, sound, level of grouping, and language, but is a fairly robust generalization. The reverse pattern, where initial or final elements in a group are *shorter* than their medial counterparts, is extremely rare, except in cases where boundary-adjacency is in complementary distribution with stress-adjacency (e.g. Yoloxóchitl Mixtec, DiCanio et al., 2022).

2.2.2 Pitch

Linguistic pitch is used for many different purposes. Relevant to grouping is the existence of *edge tones*, pitch targets or movements that occur at the edges of groups. The main acoustic correlate of perceived pitch is the fundamental frequency (f0) of a sound. Linguistic theories include edge tones because many languages tend to display extreme f0 values and/or f0 movement at the beginnings and ends of groups. These are not necessarily the most dramatic f0 discontinuities in an utterance, particularly in languages with pitch accents, but patterns are robust enough to require a theoretical description.

In American English, f0 movement tends to occur at the ends of groups corresponding to syntactic phrases (e.g. Pierrehumbert, 1980). These f0 changes do not correspond to stress or lexical tone, but to the illocutionary force or discourse function of constituents. Most researchers posit at least two levels of grouping in English that generate edge tones (e.g. Beckman and Pierrehumbert, 1986; Ladd, 1986), which means that higher-level group boundaries will tend to have more tonal targets and hence more f0 movement than lower-level boundaries, all else being equal. For instance, one analysis of the English 'continuation rise' posits a combination of edge tones at the intermediate- and intonational-phrase levels of grouping (Pierrehumbert and Hirschberg, 1990).

Similar edge-tone systems exist in a variety of languages that differ from English in the presence and nature of stress, pitch accent, and lexical tone. Bengali, for instance, differs from English in having almost entirely predictable word stress, but its intonational system

also features edge tones at two levels of grouping (Hayes and Lahiri, 1991). Tokyo Japanese lacks word stress but displays lexical (unpredictable) pitch accents in some words (McCawley, 1968). Groups corresponding to single content words with adjacent function morphemes are generally realized with an initial high tone and a final low tone, resulting in a sharp f0 rise phrase-initially. This rise is larger ('pitch reset') at the beginnings of groups that correspond to larger syntactic constituents (Pierrehumbert and Beckman, 1988). Seoul Korean plausibly lacks both stress and pitch accent; groups are delimited most consistently by a final rise in f0 (Jun, 1993; Kim, 2004). As in English, higher levels of grouping involve the agglutination of additional edge tones (Jun, 1993). Sri Lankan Malay, which lacks stress and accent, also features groups demarcated by final f0 rises (Nordhoff, 2012).

Some languages with lexical tone also tend to align more complex or dynamic f0 patterns near group boundaries, sometimes but not always involving edge tones. Thai has lexical tone and predictable stress (Tingsabadh and Abramson, 1993). Many lexical tones take a complex and dynamic form at the ends of phrases but are simplified and flattened phrase-internally (Morén and Zsiga, 2006). This is due not to edge tones, but rather to simplification of lexical contour tones everywhere *except* at the edges of groups. Bantu languages, on the other hand, generally display lexical tone but lack word stress. Many of them feature penultimate lengthening at the utterance, phrase, or word level of grouping. This lengthening is frequently accompanied by edge tones that produce more complex f0 contours on the second-to-last vowel in a group than in other prosodic contexts (Hyman, 2013).

In sum, many languages display 'extra' tonal movement at the edges of groups, compared to group-internal contexts. While the amount of f0 movement in any given context can also be affected by factors such as metrical prominence and lexical tone, edge tones are attested independently of these phenomena and can interact with them.

2.2.3 Consonantal manner

A third property dependent on grouping structure is *manner*, a set of phonetic features pertaining acoustically to the magnitude and velocity of changes in intensity. In articulatory terms, manner corresponds roughly to the degree of constriction in the vocal tract: vowels are associated with relatively open vocal tract configurations, and consonant configurations range from wide and vowel-like (approximants) to extremely narrow (obstruents).

Consonants and vowels are often longer and less vowel-like at group boundaries, shorter and more vowel-like within groups. The *initial strengthening* literature finds that consonants are longer at the beginnings of larger groups, and also that the articulatory gestures associated with their constrictions are more extreme (e.g. Byrd and Saltzman, 1998; Keating et al., 2003; Onaka et al., 2003). In English, words that begin with vowels are more likely to display initial glottal constrictions at the beginnings of larger groups (Dilley et al., 1996; Pierrehumbert and Talkin, 1992). English vowels are also more likely to occur with glottalization at the *ends* of larger groups (Redi and Shattuck-Hufnagel, 2001). And English vowels have more extreme backness and height values at the beginnings and endings of larger prosodic domains compared to smaller ones (Cho, 2005).

Phonological theory describes group-dependent manner differences in terms of *lenition* and *fortition*. While these terms are used to describe a broad and heterogeneous set of phonetic patterns, there is a 'core' set of lenition patterns seen in many language families that tend to target medial consonants at one or more levels of grouping, making them less

constricted and/or shorter than their initial counterparts [see Kirchner (1998) and Lavoie (2001) for typological surveys]. Typical lenition processes that apply internal to prosodic domains, most often between vowels or approximants, include voicing, spirantization, and tapping. These processes take sounds that would generally have lower acoustic energy at the beginnings of domains, and turn them into sounds with more acoustic energy internal to domains by lessening their degree of constriction or changing laryngeal properties. The set of languages displaying one or more of these patterns is too large to enumerate here: a small selection of studies with substantial phonetic detail includes Spanish (Hualde et al., 2011; Kingston, 2008), American English (Bouavichith and Davidson, 2013; De Jong, 2011), Gurindji (Ennever et al., 2017), Campidanese Sardinian (Katz and Pitzanti, 2019; Katz, 2021), and Iwaidja (Shaw et al., 2020). There are dozens of other languages where such patterns have been described in qualitative phonological terms.

Kingston (2008) and Katz (2016) propose that these typologically ubiquitous patterns align larger changes in intensity with group boundaries and smaller changes with non-boundaries. On this view, voicelessness and stopping are favored at group boundaries because they create larger changes in intensity between two flanking vowels or approximants, while voicing and continuancy are favored medially because they entail smaller changes from surrounding vowels or approximants. The exact nature and contexts of lenition and fortition phenomena can be quite complex, often varying with the phonetic features of the affected sounds or adjacent sounds, but they tend to occur between vowels if they occur anywhere (Kirchner, 1998; Lavoie, 2001).

In sum, a common effect of group boundaries on acoustic patterning is that consonants tend to be less vowel-like adjacent to boundaries, more vowel-like medially. The result, if consonants are adjacent to vowels, is larger changes in intensity near group boundaries and smaller changes medially. While we have presented these phenomena separately from the duration differences laid out in the preceding section, there is a fair amount of evidence that intervocalic lenition, whatever its ultimate cause, is often mediated by duration shortening (Ennever et al., 2017; Katz and Pitzanti, 2019; Cohen Priva and Gleason, 2020; DiCanio et al., 2022).

2.2.4 Summary of linguistic grouping

All of the canonical prosodic processes described above can be viewed as instantiating Gestalt grouping constraints, in particular the proximity and change principles described in Sect. 1.2. Beyond the presence and absence of prosodic boundaries, there are many other factors in language that will affect the proximity and similarity of speech sounds. But holding those other factors equal, the common prosodic patterns just outlined will tend to make adjacent sounds within a prosodic group more similar and more proximal to one another than adjacent sounds spanning a group boundary.

Because final and some initial elements within groups have a tendency to be lengthened, the onsets of two speech sounds (or syllables, words, etc.) will tend to be further apart in time when they span a group boundary than when they do not, all else being equal. This tendency should be even more pronounced when a pause occurs at a prosodic boundary. While pauses during speech can occur for many different reasons and do not always correspond to prosodic boundaries, there is a robust tendency for some prosodic boundaries to be marked by pauses (Gee and Grosjean, 1983; Ferreira, 1993; Choi, 2003, see Krivokapić, 2007 for review).

Pitch and intensity are both relevant to the similarity principle. The general idea is that keeping an acoustic parameter at a relatively steady level results in continuity, while changes in a parameter result in discontinuity or disruption. So prosodic patterns that create changes in pitch or intensity at prosodic group edges, but fail to do so internal to groups, will tend to align acoustic discontinuities with group boundaries and relative continuity with the lack of boundaries, all else being equal. Just as with duration, there are many linguistic and extra-linguistic factors that affect the pitch and intensity of speech sounds.[2] But the presence of edge tones associated with material at the endings and/or beginnings of prosodic groups will create a tendency for greater pitch movement in the general vicinity of group boundaries, all else being equal. Pitch-range resetting at the beginnings of prosodic groups will have a similar effect. And prosodically-conditioned changes in the sonority or intensity of consonants will tend to make them more similar to flanking vowels or approximants internal to a group (*lenition*), and less similar to such sounds at a group boundary (*fortition*).

There are hundreds of prosodically-conditioned sound patterns documented in phonetics and phonology, and this section hasn't covered all of them. But the patterns discussed here are quite robust cross-linguistically and a substantial chunk of all described prosodic processes could be categorized as one of these general types. Each of these patterns also instantiates a Gestalt grouping principle of the type that is said to be relevant to musical grouping. In the next section, we survey the Gestalt view of musical grouping.

3 Grouping in music

While musical grouping has not been studied as thoroughly as its linguistic counterpart, there is broad agreement as to how it works in general terms (details, of course, differ between theories). Lerdahl and Jackendoff (1983), in their *Generative Theory of Tonal Music* (henceforth *GTTM*), lay out a theory of musical grouping in great detail and subsequent work provides a fair bit of empirical support for their description. Other models (e.g. Cambouropoulos, 2001; Narmour, 1990) differ in details and orientation but tend to agree on the broad characteristics of grouping. The motivation for grouping in music, just as in language, pertains to the relationship between sound structure and 'syntactic' structure (equated here with the combinatorics of musical harmony). In particular, *GTTM* argues that: (1) experienced listeners tend to parse musical events into certain kinds of constituents on the basis of auditory features; and (2) those constituents are relevant to interpreting the harmonic, rhythmic, and/or thematic information within a piece, as well as describing the representation of pieces of music in memory. Harmony is addressed separately in *GTTM* as *prolongational reduction*. The theory is couched in somewhat different terms than the linguistic theory of prosody and based on different kinds of evidence. But the general similarity of the resulting structures, as Lerdahl and Jackendoff note, is striking.

3.1 General properties of musical grouping structure

Grouping in music is sometimes investigated experimentally, through implicit or explicit tasks examining the 'chunking' of musical events in memory. In *GTTM*, it is generally inferred on the basis of a skilled listener's intuitions (just as in some phonological

[2] In particular, we are still ignoring metrical prominence entirely, which can have a dramatic effect on pitch, duration, and intensity in some languages, and would thus be expected to interfere with the alignment between prosodic boundaries and acoustic changes/discontinuity. Section 4.2 revisits this issue.

research). Those intuitions in turn are based on auditory properties described in Sect. 3.2. Before turning to those properties, however, I outline the general structure of musical groups. As in language, grouping involves sets of constituents nested hierarchically inside larger constituents. Figure 3 shows a possible grouping structure for the traditional folk song 'Turkey in the Straw'. As in Fig. 2, the grid notation is shown above the example and *GTTM* notation below; the two are formally equivalent. One reason for choosing this song is that it has no standard set of lyrics and is usually an instrumental piece, so we can be certain that constituency here is inherently musical rather than inherited from linguistic representations. That said, less musically inclined readers may find it easier to imagine a familiar song with words and note how linguistic constituents map to musical events.

Just as in the linguistic example, this grouping structure is not uniquely determined by the properties of the tune. Other listeners could disagree about the exact location of grouping boundaries. And it is possible that different performances of this tune could evoke different grouping structures. But all possible grouping structures tend to share some basic properties. *GTTM* states these properties as Grouping Well-Formedness Rules. Several of them are instructive for comparative purposes. Occurring musical events are exhaustively partitioned into groups. If a group *X* contains some of the elements in another group *Y*, then either *X* or *Y* contains all of the elements in the other group (with one exception involving transformation rules). Every piece coincides with a single group that contains all other groups. And there is a tendency for groups to contain exactly two smaller groups. These representational properties of musical grouping are inferred from intuitions about constituency and (indirectly) from harmonic interpretation. The reader may notice that they are exactly the same as the properties of linguistic grouping structure discussed in Sect. 2.1.

The claim of *GTTM* is that the type of groups picked out here are also the constituents relevant to a harmonic or motivic analysis of the excerpt. For instance, all of the groups in Fig. 3 end on local *consonances*, notes contained in the local harmony implied by the piece. At a higher level of structure those notes outline a tonic triad progressing to a dominant chord in the first half of the excerpt, then again a tonic triad progressing to a V-I cadence to close the excerpt. The explicit claim of *GTTM*, and a tacit methodological assumption of most traditional music theory, is that these types of melodic and harmonic patterns are central to the tonal interpretation of a piece and that they bear a systematic relationship to groups picked out on the basis of Gestalt principles. For instance, this

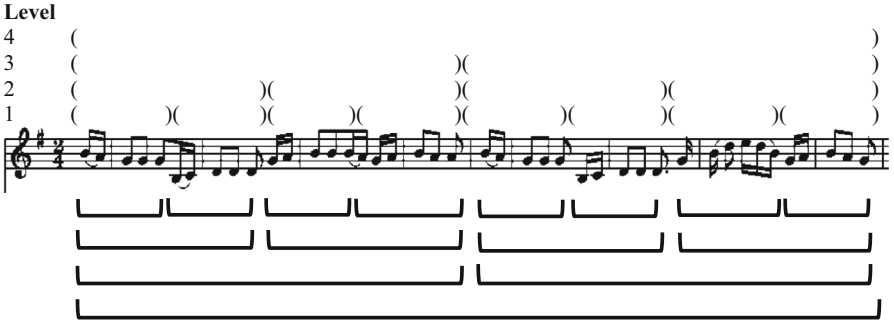

Fig. 3 A possible grouping structure for the folk song *Turkey in the Straw*. The same grouping structure is shown in bracket notation (above the music) and Lerdahl and Jackendoff's (1983) grouping notation (below the music).

excerpt of 'Turkey in the Straw' forms a particular type of phrase known in musicology as an antecedent-consequent period, much like the opening of Mozart's *Piano Sonata No. 11* in *A*, K. 331 (Lerdahl and Jackendoff, 1983, p. 32), and hundreds of other examples from the history of Western Tonal Music. The defining properties of a period include both harmonic criteria (the final tonal closure is more conclusive than the medial one at the halfway point) and rhythmic criteria (that medial half-cadence *is* in fact near the halfway point of an 8-measure unit). And the claim of *GTTM* is that grouping structure comprises the rhythmic domains from which we select harmonic events (half-cadence, perfect authentic cadence) that identify 'Turkey in the Straw' as being of the class 'antecedent-consequent period', and identify the Mozart theme as being of the same class. Crucially, the constituent structure relevant to this type of harmonic and thematic analysis is the same structure that Gestalt grouping principles converge on. We turn to those principles next.

3.2 The acoustics of musical grouping

Whereas the mapping between sound patterns and grouping in language tends to be approached as a function from groups to sounds, *GTTM* approaches musical grouping in the opposite direction, as a function from sounds to inferred grouping structures. Despite this reversal, the actual content of the two functions is quite similar. Pearce (2008) offers a detailed and thorough review of grouping theories and empirical results, and some of the review here is based on that discussion. In Sect. 3.2.1, the Gestalt principles are illustrated with regard to the Western song in Fig. 3, and I review evidence for Gestalt grouping in Western music for Western listeners. Section 3.2.2 reviews numerous other examples of Gestalt grouping in musical genres from around the world.

3.2.1 Gestalt principles in Western music

We saw in Sect. 2.2 that linguistic group boundaries often coincide with longer events and larger changes in f0 and intensity. The same is true in music. *GTTM* posits violable constraints calling for moments of auditory disruption in the musical surface to be aligned with group boundaries. Auditory 'disruption' here is conceived of as a change in some auditory parameter, where that parameter is comparatively steady before and after the change, or as temporal discontinuity in the sense of periods of time with no event onsets in them.

The proximity principle entails that notes whose onsets are less temporally distant from one another are less likely to be perceived as spanning a group boundary. In Fig. 3, for instance, the proximity principle predicts the boundary between the sixth and seventh groups, which falls after a note longer than surrounding ones. Note that 'Turkey in the Straw' was chosen in part because it doesn't involve many long notes or rests. The proximity rule tends to dominate musical grouping perception, just as overt pauses strongly indicate high-level prosodic domains (or planning difficulties) in speech. In the absence of clear and unambiguous duration cues, other grouping principles emerge more clearly. Again, it will likely be easier for some readers to understand this principle with reference to familiar songs with lyrics: the proximity principle means that large linguistic units such as phrases or sentences are often separated by pauses or long notes. They also tend to correspond to harmonically and melodically coherent units.

The similarity principle entails that notes whose pitch, timbre, articulation, or intensity (among other properties) are less distinct from one another are less likely to be perceived as

spanning a group boundary. This principle predicts the boundaries in Fig. 3 between the first, second, and third groups, which occur at local maxima for pitch change. As notated in this particular transcription of the tune, the slur marks underneath notes suggest that there could be a change in articulation between the first two notes in each group and the following notes, in which case the similarity principle would also call for grouping boundaries at those junctures. I haven't included those groups in Fig. 3 both for reasons of visual clarity and because it's not clear what the basis is for the slurring notated in this transcription. For readers who find it easier to reason about familiar songs with words, the similarity principle would suggest that pitch movements at the beginnings and endings of large linguistic constituents are often smaller than pitch movements between the ending of one phrase and the beginning of the next. This can be observed, for instance, in the Happy Birthday song. This song also includes pitch leaps *internal* to linguistic constituents, which may provide evidence for smaller musical groups.

The similarity principle also predicts group boundaries on the basis of discontinuities in intensity and timbre, which would require an actual performance in order to judge, rather than a written representation (Western music notation can optionally include aspects of loudness and differences in timbre, but much of the variation along these dimensions is left to the discretion of the performer). Though not all of the similarity principles can be straightforwardly illustrated with a written representation at this level of abstraction, the overarching idea should be clear: we predict that changes in intensity or timbre will tend to be aligned with group boundaries, and that such changes will themselves encourage the perception of a group boundary.

None of the grouping theories discussed here claims that group boundaries are always marked by specific acoustic changes, or that acoustic changes always mark a group boundary. In actual music, different Gestalt principles may conflict with one another, suggesting different locations for group boundaries, or there may not be strong acoustic discontinuities at all. In all of these cases the perceived grouping will be somewhat ambiguous. 'America the Beautiful', for instance, displays a number of grouping cues in conflicting positions, none of which seem to clearly align with linguistic constituents in the lyrics. The most we can say about such tunes from the perspective of grouping theory is that they're ambiguous (and that the music doesn't converge on the same group boundaries as the lyrics). In general, the nature of the *GTTM* approach to grouping is that specific auditory discontinuities are associated probabilistically with the presence of a group boundary, where the magnitude of disruption and the weighting of different acoustic cues factor into the final percept of groups.

Basic grouping principles have been robustly confirmed for Western musicians and non-musicians through explicit grouping tasks (Deliège, 1987; Peretz, 1989) and implicit tasks examining the influence of grouping on memory (Deutsch, 1980; Dowling, 1973; Tillmann and McAdams, 2004). There is evidence that infants use the proximity principle (Jusczyk and Krumhansl, 1993; Krumhansl and Jusczyk, 1990). And higher-level group boundaries have a cumulative, hierarchical effect on production, perception, or recall (Stoffer, 1985; Todd, 1985; Large et al., 1995).

3.2.2 Gestalt principles and grouping across cultures

Given the generality and psychophysical nature of Gestalt principles, musical grouping ought to show gross similarities across cultures with respect to proximity and similarity. There has been a limited amount of work on Gestalt principles in non-Western music

(summarized below). However, ethnomusicological descriptions of specific musical genres often allow us to infer the presence and nature of grouping principles as well. In this section, I review a number of studies of specific non-Western musical genres that illustrate proximity- and similarity-based grouping.

Review articles on universals in music frequently mention Gestalt principles of grouping as one such putative universal (e.g. Harwood, 1976; Higgins, 2006; Meyer, 1991; Stevens and Byron, 2009; Trehub, 2000). These assertions tend to be based on evidence from the infant literature (e.g. Jusczyk and Krumhansl, 1993; Thorpe et al., 1988), or simply on impressionistic observations from experienced musicologists. Few of these papers contain any examples of or details about grouping in non-Western genres. Nonetheless, a number of detailed studies exist that either implicitly or explicitly demonstrate Gestalt grouping at work in the performance, composition, or perception of world musics.

Ayari and McAdams (2003) present a detailed study of Arabic Taqsim music, an improvised instrumental form, and its perception by skilled musicians from Arab and Western cultures. In terms of performance, their transcriptions clearly show that small 'segments' introducing particular tonal material at a local level, as well as higher-level formal sections associated with different tonal collections, tend to be marked by disruptions in at least pitch-level, intensity, and proximity (long notes or pauses). In terms of perception, only the segmentations of Arab musicians reflect culture-specific tonal properties, but both Arab *and* Western musicians' segmentations show an influence of Gestalt proximity and similarity. Mungan et al. (2017) study a related form, Turkish Makam music, but with both trained and untrained listeners and more tightly controlled musical materials that eliminate most boundary cues apart from proximity and melodic contour. They find a high degree of concordance between locations of disruption in Gestalt proximity and the segmentations of all groups of listeners: Turkish musicians, Turkish non-musicians, and Western non-musicians.

Several studies in other non-Western genres show that Western listeners converge on the grouping judgments of 'native' listeners to those genres in ways that reflect Gestalt proximity. These studies generally also demonstrate, at least implicitly, that Gestalt proximity is a factor in the *performance* of the music in question. Popescu et al. (2021) give such an illustration for North Indian sitar alap, using naïve Western listeners and Indian experts; they leave open the possibility that Gestalt similarity also plays a role in perception and production, but only provide direct evidence for proximity. Nan et al. (2009) report effects of proximity-based grouping in Western classical and traditional Chinese music, across German and Chinese listeners; Schellenberg (1996) shows that pitch similarity also helps predict segmentation of Chinese pentatonic folk songs for Chinese and Western listeners.

Temperley (2000) argues on the basis of existing ethnomusicological work that (mostly Ewe) West African music clearly involves grouping on the basis of proximity, and may involve various similarity principles as well (cf. Agawu, 1990 on the alignment of 'structural importance' and rhythmic prominence/duration in Ewe songs). Pasciak (2017) proposes a particular tonal analysis of Japanese Edo-period koto music, as well as a critical review of earlier formal literature on the genre. Many of his detailed examples focus on transitions within a piece between different tonal collections identified as *tetrachords*. The examples also show implicitly that such tonal areas have a strong tendency to be separated by proximity- and pitch-similarity-based grouping principles (see, e.g., figures 15 and 32 from Pasciak, 2017). Hughes (1988) offers a generative theory of Gendhing Lampha music from central Java (a genre of gamelan music). The analyses here are dense and complex;

one clear aspect, however, is that the music is organized around quaternary metrical constituents called *gatra* that play an important role in structuring melodies. In some (though not all) pieces, successive gatra are separated by pitch leaps in one or more pitched instruments. Groups of several gatra are organized into *gongam*, high-level formal sections; some of these gongam are marked by final pauses or long notes.

I have tried to focus so far on non-vocal music, because it is in some sense a stronger test than vocal music of the idea that music has inherent grouping principles: if we find evidence for Gestalt grouping in vocal music, one may wonder whether it is 'inherited' from language. That said, a substantial portion of all musics in the world involve singing, the existence of Gestalt grouping principles has been robustly documented in sung music across cultures, and the idea that music only inherits its grouping from linguistic constituency is not actually a viable theory when we attend to details. Many studies show that linguistic constituents in the text sometimes or always match up with musical rhythmic units that are separated by pauses or long inter-onset intervals (IOIs). This is true in English children's songs and art music (Halle, 2004), American hip-hop (Horn, 2010; Katz, 2015), French traditional song (Dell, 2015), Hausa rajaz (Hayes and Schuh, 2019), and Tashlhiyt Berber song (Dell and Elmedlaoui, 2002). Stock's (1993) transcription of a Kalasha praise song from Northwest Pakistan shows 'line' and 'half-line' constituent levels separated by Gestalt proximity and changes in intensity. Blacking's (1970) description of Venda songs from South Africa shows that melodic cells analyzed as the fundamental building blocks of songs are generally separated by 'tones stressed by duration and meter', as well as disruptions in pitch similarity. Kaliakatsos-Papakostas et al. (2014) apply the *GTTM* grouping theory to polyphonic Balkan odd-metered music from Epirus, using the Gestalt principles to single out constituents relevant to harmonic and rhythmic analyses. McPherson and Ryan (2018) focus mainly on lexical tone mappings in Tomma So (Dogon) women's songs and do not examine grouping. The scores they include in an appendix, however, make it clear that high-level musical phrases corresponding to large linguistic constituents are set apart from one another by both long IOIs and, frequently, pitch leaps larger than the preceding or following intervals.

While all of these genres show Gestalt-based grouping at relatively high levels (often referred to as 'lines'), in some of them duration patterns at lower levels are more closely associated with linguistic *stress* or *weight* than grouping (see Sect. 4.2). Examples of Gestalt-based grouping at both higher and lower levels, however, can be seen in Turpin's (2007, 2011, 2017) detailed analyses of several song genres from the Arandic cultures of central Australia. For instance, Turpin (2007) shows that the Akwelye genre of women's songs aligns every phonological word with a rhythmic 'cell', and transcriptions make it clear that those cells are separated by long IOIs. At higher levels, groups of such cells align with conventionalized melodic sequences, and those higher-level sequences add pitch-similarity-based grouping cues to the proximity cues found at all levels.

The only example I've run into in the course of this survey that shows no clear evidence for proximity or pitch-similarity as a factor in grouping is Ekwueme's (1975) transcription of an Igbo musketeers' song. Interestingly, the major musical constituents in this song, which show no particular tendency to be separated by long IOIs or pitch leaps, are sung in alternation by a soloist and chorus. This means that, almost by definition, the phrases here are separated by disruptions in timbre, texture, and/or intensity along Gestalt lines.

To summarize, the edges of performed, composed, and perceived musical groups in many genres across the world tend to be marked by disruptions in pitch and intensity and by temporal disjuncture. 'Disruption' here refers to a change in some acoustic parameter between two events where the surrounding events do not show as much of a change. These generalizations are plausibly a product of domain-general Gestalt principles. And they bear

an obvious resemblance to the marking of boundaries in linguistic grouping. In the final section, I compare the two systems in more detail.

4 Parallels and non-parallels between musical and linguistic grouping

The similarities between musical and linguistic grouping outlined in the preceding sections are relatively clear: in both domains, constituents that are relevant to some external domain (syntax, pragmatics, memory, harmonic or semantic interpretation) display relative acoustic continuity internally and disruption at their edges. In both domains, the relevant constituents appear to be hierarchically nested. The types of evidence used to support these conclusions, however, are somewhat different in the two domains, and so are the resulting theories. Theories of linguistic grouping are based on the distribution of sounds in speech production, which are taken to be a product of structural factors. Theories of musical grouping are based on intuitions (including systematic experimental investigation of intuitions) or chunking in memory, and take groups to be a product of auditory continuity and disruption along broadly Gestalt lines. The precise details of which acoustic parameters contribute to grouping and how they do so also differ in the two domains. In this section, I examine these differences in greater detail and attempt to draw some general conclusions about the two systems.

4.1 Directionality, production, and perception

In generative linguistics, the *grammar* is taken to be a body of implicit knowledge that includes the principles governing linguistic structure-building. This means that while the grammar influences all aspects of linguistic behavior, it is described not in terms of algorithms that drive specific production and processing activities, but in terms of the 'final state' of information that algorithms for specific processes may draw upon. One common model is a function from arrays of *lexical items* to the set of well-formed utterances in the language in question, including both their sound patterns and semantic interpretations. We treat lexical items here as tuplets stored in long-term memory that capture arbitrary associations between sound, meaning, and syntactic features. In reality, the structure of the lexicon is probably more complex than this: 'atomic' units of stored meaning, stored syntactic features, and stored sounds may fail to align with one another in various ways (Anderson, 1982; Beard, 1986; Halle and Marantz, 1993), to the extent that the traditional notion of *morpheme* may not even be well-grounded. But such facts will not make any difference to the relatively 'high-level' points we make here about the difference between language and music: a more realistic theory will complicate our understanding of how and for which units the arbitrariness of sound-to-structure correspondence is computed, but not the existence of such arbitrariness.

If linguistic grammar is conceived of as a function from lexical items to well-formed linguistic structures, then there are various intermediate steps, including a function from syntactic structure to prosodic grouping and one from prosodic groups of lexical items to the phonetic realization of those items. The *GTTM* model of grouping is different because the entire form of the musical grammar in this theory is different. In *GTTM*, the grammar is also a final-state theory of the information that specific musical processes may call upon. But it is described as a function from the auditory surface of musical pieces to a set of metrical, grouping, and harmonic analyses of those pieces. This is in some ways the opposite of its linguistic counterpart.

One question that immediately arises is whether the directionality in these descriptions is actually necessary. These are fundamentally descriptions at what Marr (1982) refers to as the computational level, and as such they are not models of actual behavior. So the interpretation of directionality in these theories need not be in terms of processing vs. production behavior, but instead should be viewed in terms of information flow, predictability, and redundancy.

In linguistics, it is generally believed that there is information loss in the mapping from syntactic structures to prosodic ones: grouping is insensitive to some of the distinctions that matter for syntactic constituency. Although theories differ, examples might include distinctions between syntactic arguments and modifiers, or between structures with more or fewer levels of embedding under function morphemes. In other words, most theories predict that more than one syntactic structure can be mapped to the same grouping structure. That said, the more recent 'recursive' approaches to prosody discussed in Sect. 2.1 entail less information loss between syntactic and grouping structures, possibly none.

At the level of linguistic behavior, it is clear that listeners recover the syntactic structure of an utterance partly based on perceived grouping. Acoustic reflexes of grouping influence morphological and syntactic processing in precisely the ways that Gestalt grouping principles would predict. At lower levels of syntactic constituency, such as the morpheme or word, there is abundant experimental evidence bearing on the segmentation of lexical items. As a starting point, the detection of such constituents for infant and adult listeners depends in part on transitional probabilities between speech sounds (e.g. Mattys and Jusczyk, 2001; McQueen, 1998) and between syllables (e.g. Saffran et al., 1996a, 1996b): all else being equal, lower-probability transitions are more likely to be inferred to span a group boundary than higher-probability transitions. The relevant aspect of this research for our current purposes is evidence, from artificial and natural languages and from listeners of all ages, that all of the acoustic cues to grouping discussed in section 2.2 can reinforce or conflict with statistical cues and affect segmentation (Bagou et al., 2002; Christophe et al., 2003; Katz and Fricke, 2018; Kim, 2004; Millotte et al., 2011; Nakatani and Dukes, 1977; Saffran et al., 1996b).

At higher levels of syntactic constituency, cues to prosodic grouping help listeners adjudicate between competing syntactic analyses of similar strings of words (e.g. Price et al., 1991; Schafer et al., 2000). The precise nature of the cues involved and the details of online processing are a matter of debate (see Carlson et al., 2001), but the general fact that a description of sentence processing must make reference to grouping is not in doubt.

So while theories of linguistic grammar generally map from syntactic structure through grouping structure to sound patterns, there is no doubt that listeners 'reverse-engineer' this mapping to recover constituency from the acoustic stream. Beyond this, several theories within the broad tent of generative linguistics propose a more symmetrical relationship between syntactic structure and grouping. Richards (2010, 2016) argues that the description of syntactic computations across languages must make reference to language-specific principles of grouping. Jackendoff (2002) proposes that syntactic and grouping structures are independently generated and relations between them are enforced by correspondence constraints. Steedman (2000) argues that grouping directly reflects the constituency of information structure, with highly flexible relationships between each of these domains and syntactic structure. Taken together, these considerations suggest that the directionality of syntax-grouping correspondence in traditional linguistic theories may be largely a matter of convenience, orientation, and/or methodology, rather than reflecting deep properties of information flow within grammar.

If the relations between syntax and grouping and between grouping and sound patterns in language are non-directional, is the same true of music? In musical grouping theories, including *GTTM*, there is a fully transparent relationship between auditory disruption and grouping, and there is no reason that the mapping could not be reversed to derive probabilities of various types of disruption from grouping structure. Just as it is true that a rest, leap in pitch register, or change in dynamics has a non-null probability of being perceived as marking a group boundary, any given group boundary has a non-null probability of being marked by one of these acoustic discontinuities. The precise details of which types of disruption are most frequent at which group boundaries may be an interesting area of study. At a first pass, *GTTM* notes that higher-level group boundaries tend to be associated with larger acoustic disruptions, and with more different kinds of acoustic disruption; several of the non-Western genres reviewed in Sect. 3.2.2 suggest something similar.

In terms of grouping and syntax, *GTTM* entails that more than one grouping structure can map to the same syntactic (reductional) structure, but the converse is equally true. And while the majority of the *GTTM* principles concern the mapping from musical surface to grouping and from grouping to syntax, the authors are careful to note that information from the syntactic reduction components is also necessary to adequately describe the grouping system. One major question, then, is whether syntactic constituents in music could be described as generating a probability distribution over possible grouping implementations, as the relationship is generally described in language. If there is information loss from grouping to syntactic structure, then such a model would be impossible and the directionality of the *GTTM* algorithm would be necessary. Katz and Pesetsky (2011) show that the *GTTM* algorithm governing the relation between time-span reduction (reflecting grouping constituency) and prolongational reduction (reflecting syntactic relations) can in fact be cast in non-directional terms with little or no loss of information.

The overall conclusion here, then, is that neither traditional linguistic theories of the syntax-grouping interface nor the *GTTM* theory of the musical syntax-grouping interface actually need the directionality that is built into them. This difference instead seems to be a matter of convenience or orientation, as suggested above for language. However, the claim that the relationship between grouping and harmonic syntax is non-directional, following Katz and Pesetsky (2011), does not answer the independent question of where musical syntactic structure comes from if it is not derived from grouping. And this question gets at a difference between *GTTM* and linguistic theories that really *is* profound and substantive.

GTTM differs from most linguistic approaches in that it takes 'the set of well-formed musical pieces' to be either a given or an incoherent concept. As such, the theory describes how pieces of music are assigned structure, but does not attempt to describe why some pieces are more or less likely to exist than others. This difference is independent from the syntax-grouping interface discussed above: it could be the case that the interface is identical in language and music, but whether and where the model expresses generativity (in the sense of defining the difference between well-formed and ill-formed pieces) is different. In other words, we could have a theory just like *GTTM* but where the relationship between syntactic and rhythmic components is expressed non-directionally, with global correspondence constraints.

That said, many other approaches to musical harmonic syntax differ from *GTTM* with regard to generativity (e.g. Steedman, 1984; Johnson-Laird, 1991; Rohrmeier, 2011, 2020a; Katz, 2017). Each of these approaches proposes a generative syntactic component that defines all and only the well-formed harmonic progressions in some genre. This is based on a mix of arguments from the existence and non-existence of various harmonic sequences and the tonal interpretation of harmonic sequences when they occur.

Taken together, this research makes a strong case that generative syntax is necessary in music as it is in language, independently of rhythmic grouping.

Rohrmeier (2020b) sketches an attempt to connect generative theories of musical harmonic syntax to grouping structure, but there is not a lot of other work in this vein. More generally, evidence that some independently motivated notion of 'syntactic constituent' in music corresponds in a systematic way to musical grouping is much patchier than for other generalizations discussed here. This is in part because syntactic constituency is less clear in music than it is in language, because musical grouping is rarely approached from the perspective of syntactic structure, and because there is far less work on musical syntax in general than on any other theoretical area discussed in this paper.

There is a body of research showing (sometimes implicitly) that performers tend to elongate the ends of major harmonic sections (e.g. Repp, 1992, 1998; Todd, 1985) in such a manner that the resulting grouping structure is isomorphic to major harmonic (syntactic) constituents. Many of the ethnomusicological descriptions reviewed in Sect. 3.2.2 posit constituents referred to as 'phrases' on the basis of what appear to me to be tonal, motivic, and/or thematic criteria, and those phrases tend to be aligned with rhythmic groups as characterized by Gestalt principles. Beyond this, there is not much literature explicitly investigating the interaction between harmonic and rhythmic constituency in music. The most substantial collection of relevant analyses may well be those contained in *GTTM* and Lerdahl's (2001) extension of the theory. These dozens of detailed analyses show that, in the default case, the domains relevant to computing harmonic dependencies (prolongational reduction) are *the same* domains relevant to computing rhythmic prominence (time-span reduction), just as in the simplified account of antecedent-consequent period structure in Sect. 3.1. This is despite the fact that the authors frequently choose examples meant to illustrate the complexity of the mapping (in such cases, there are limited degrees of mismatch between the two types of constituent).

It should be understood that the *GTTM* prolongational theory and the generative theories above that I refer to as syntax are not universally accepted even for Western tonal music, and there are questions as to whether harmony is appropriate for syntactic analysis (some genres of music have weak or nonexistent notions of harmony). That said, within harmonic traditions, the balance of evidence suggests that major structural markers (e.g. cadences, tonic returns, the beginnings of major harmonic sections) tend to be realized in ways that set them apart from surrounding material in terms of grouping. This is such a basic aspect of such genres that it is more likely to be presupposed than actively investigated.

To summarize, the differences between *GTTM* and standard linguistic theory in characterizing syntax-grouping correspondence do not necessarily reflect any deep differences in information-flow between the two underlying cognitive systems. In language, there is evidence that the mapping from syntax to grouping is less directional than appreciated in the early stages of prosodic theory. And in music, many or all of the mappings from sounds to groups and groups to syntactic constituents are also non-directional. The difference in presentation has more to do with *GTTM*'s overarching goals, which are different from most linguistic approaches, and with the fact that grouping in music is easier to intuit (and more widely agreed upon) than syntactic structure. Other approaches to musical syntax are more closely aligned with linguistic models, and there is some work illustrating how these syntactic models might line up with grouping, but this is a generally sparse area of research that requires more investigation.

4.2 Acoustics, prominence, and underlying representations

The preceding sections have argued for computational-level similarities between musical and linguistic grouping: implicit knowledge about the structure of well-formed groups and their relationship to auditory properties overlaps substantially between music and language. Similarities in the acoustic reflexes of musical and linguistic grouping are striking enough to raise the question of whether listeners could use the same grouping algorithm to recover constituency from the auditory stream in the two domains. I suggest here that the answer is probably not, and the reasons why tell us something interesting about basic 'design principles' of music and language.

While theorists disagree on precisely which aspects of musical events are crucial to computing combinatoric (syntactic) dependencies, all theories share one broad commonality: the categories relevant to syntax severely underdetermine actual acoustic realizations. For instance, in a theory where chords are the basic building blocks of syntax, there is an infinite number of ways that a given chord can be performed: more or fewer notes, longer or shorter duration, higher or lower pitch and intensity, faster or slower attack, etc. As all of these acoustic parameters are freely varied by composers and performers to demarcate groups, grouping algorithms do fairly well by simply scanning a representation of musical notes and locating acoustic discontinuities. For instance, Thom et al. (2002) show that a variety of simple grouping algorithms based on very few acoustic parameters produce good agreement with human annotators, even using a measure that ignores a large portion of the agreeing cases (where neither humans nor models infer a group boundary).

To simplify somewhat, a listener who hears an acoustic discontinuity in music can be relatively confident that it marks a group boundary. In language, on the other hand, there are a number of sources for auditory discontinuity besides grouping *per se*. Perhaps the most obvious is metrical prominence, generally referred to as *stress* at the level of words and *accent* at higher levels. The presence and absence, position, and acoustic implementation of stress and accent all vary across languages. But the most frequent acoustic parameters of prominence are precisely those used for marking group boundaries: pitch extrema, changes in intensity (including changes in specific frequency bands), and longer duration [see Ortega-Llebaria and Prieto (2010) for a concise review]. This means that acoustic discontinuities in the speech stream may correspond to prosodic group boundaries or they may correspond to a prominent syllable (among other things). So at a bare minimum, any grouping algorithm for a language will need to be supplemented with language-specific principles that help relativize acoustic disruption to the metrical prominence of the material being parsed.

Despite the fact that boundary phenomena and stress phenomena in language are generally understood to affect 'the same' acoustic parameters, there have also been suggestions that the two structural properties may differ in the fine details of specific phonetic parameters and how they relate to one another. For instance, Wagner and McAuliffe (2019) show that phrasal prominence in English increases both intensity and duration, while phrase-finality boosts duration and *decreases* intensity. Wagner (2022) extends this result to word-level grouping and prominence in English, and shows that English listeners use their implicit knowledge of the co-variation between these two acoustic parameters to disambiguate grouping and prominence. Cho (2005) also picks out several differences in how prominence vs. grouping affect the articulation of English vowels. Katz and Pitzanti (2019) show that prosodic grouping in Campidanese Sardinian affects consonant intensity indirectly in ways that are entirely predictable from consonantal duration, but that stress

results in a different and less predictable relationship between duration and intensity. So it is possible that even if prosodic prominence and prosodic grouping generally affect the same acoustic parameters, listeners may have subtle ways of 'factoring out' these influences into orthogonal perceptual dimensions. In any case, though, this would still be a substantive difference from music.

Why doesn't this issue arise as much in music? One reason is that metrical prominence in music tends to be highly regular. While patterns of metrical prominence must be inferred from the acoustic stream, once a local metrical pattern is established it is unlikely to change during the course of a musical piece. As such, there is less need to mark metrical prominence with acoustic changes [though there is some tendency to do so, e.g. Palmer and Krumhansl (1990) on the occurrence of musical events by level of metrical prominence]. In language, on the other hand, metrical prominence is less predictable in the general case. Many languages have unpredictable stress placement within a word [see van der Hulst and Gordon (2020) for a general review of stress]. Even in languages where stress has been described as fully predictable and alternating in a regular pattern, closer inspection often reveals that the morphological composition of a word can introduce departures from regularity [see Baker (2014) for an overview of Australian languages]. Regardless of the level of metrical regularity within words, the fact that different words contain different numbers of syllables in and of itself entails that metrical regularity will be weakened or absent at the level of phrases and utterances. As such, languages with stress and/or accent virtually always mark its location using one of the acoustic parameters discussed here. Or at the very least, if a language didn't mark prominence in one of these ways, it is unlikely that linguists (or infants learning the language) would notice the prominence.

Another reason why acoustic discontinuity doesn't necessarily entail group boundaries in language pertains to a basic property of linguistic sound systems: many languages use duration, pitch, and intensity to mark contrasts in lexical meaning, that is, these cues figure in phonological features. In English, for example, intensity is one of the most obvious differences between obstruents like [k], [b], and [z] and sonorants like [m], [l], and [w] (Ladefoged and Johnson, 2011). Duration is a strong cue listeners use to discriminate voiced and voiceless fricatives (Cole and Cooper, 1975). And the perception of voicing contrasts for consonants is affected by f0 values at the beginning of a following vowel (Haggard et al., 1970). Another way of putting this is that in English, an abrupt drop in intensity, raised f0, or long duration of noise could just as easily be caused by a voiceless fricative as a group boundary. This is despite the fact that English duration and pitch are not generally considered to be primary dimensions of contrast; in other languages, they clearly are primary and would be expected to play an even larger role in discriminating speech sounds.

This means that the amount of acoustic disruption or change relevant to inferring group boundaries must be relativized not only to the metrical prominence of the linguistic material in question but also to its segmental makeup. These two properties suffice to make the mapping between acoustics and grouping a fair bit more complex in language than in music. And they both stem ultimately from one of the most profound differences between music and language: the presence of a lexicon. Speakers possess implicit knowledge about the meaningful parts (morphemes) of their languages, involving at least arbitrary pairings of sounds and meanings in long-term memory. The sounds arbitrarily corresponding to any set of morpho-syntactic features in any particular language usually have some internal temporal structure: *cat* is one syllable long in English but *feline* is two; *cat* is pronounced with two tongue body gestures, one rising to create a constriction at the velum and another associated with a low front vowel, and these two gestures occur in a fixed order. This

means that in addition to grouping meaningful units into words, phrases, and sentences, human languages also group meaningless sound features into those meaningful units. The property is referred to as *duality of patterning*; it has been characterized as a basic 'design feature' of human language (Hockett, 1960; see Ladd, 2012 for an overview and some complications), and something that sets humans apart from other animals. If nothing else, music may show that duality of patterning is not necessary to generate extremely complex symbolic systems that unfold in time.

In sum, while the relationship between acoustics and grouping is similar in music and language, the ways in which this relationship guides processing must be somewhat different in the two domains. The basic building blocks of linguistic syntax are themselves temporally complex with regard to the number and nature of the speech sounds they contain, properties which are memorized in the lexicon. This duality of patterning means that there are more independent factors contributing to acoustic continuity and disruption in language than there are in music. The question then arises: why does language display a rich lexicon and duality of patterning while music does not? There is no firm answer, but there is a common intuition that a rich lexicon is necessary to express meaning with any degree of specificity, and that recombination of meaningless elements is necessary for a lexicon of any substantial size. On this view, some rather intricate and complex differences between music and language can be traced to the lexicon and, ultimately, differences in communicative function.

4.3 Conclusion

There are several areas of similarity between grouping in music and language. With regard to acoustics, both domains make use of something like Gestalt principles of proximity and similarity. With regard to grouping structure itself, both domains involve hierarchically nested constituents. And with regard to information-exchange with other cognitive systems, both display a systematic (if noisy) correspondence with syntactic or semantic constituency. The final question I ask here is *why* such similarities exist, and what they mean for our theories of music, language, and cognition more generally.

Given the similarities sketched above, can we say that grouping principles are 'the same' in music and language? There is a sense in which they are and a sense in which they aren't. The basic principles that guide grouping in the two domains are based on the same types of information and may ultimately be rooted in properties of the environment in which human perception occurs. That said, how the principles are *deployed* in the two domains is rather different. We noted in Sect. 4.2 that the likelihood that any given acoustic disruption marks a group boundary in language must be compared (possibly through the medium of correlations with other acoustic parameters) to the likelihood that it marks something else, such as a distinction in metrical prominence or segmental features. This is less likely to matter in music.

While the *form* of grouping principles may be 'given' by general Gestalt principles, learning the grouping conventions of any genre of music or language necessarily involves assigning different weights to different acoustic cues. These weights differ quite a bit between languages, and there is no reason they shouldn't vary between genres of music as well. So a second sense in which musical and linguistic grouping might be 'the same' is if the weighting of cues in one domain affects the weighting of cues in the other. Iverson et al. (2008) argue that linguistic grouping affects non-linguistic auditory grouping in precisely this way, based on claims about grouping and acoustics in Japanese and English.

Some subsequent research converges on the finding that language experience affects non-linguistic grouping (Bhatara et al., 2016; Molnar et al., 2016), although these effects may not be robust across different stimuli and tasks (Frost et al., 2017; Langus et al., 2016), and the judgments elicited in such studies arguably confound grouping and prominence (Wagner, 2022). The majority of these experiments concern only repeating binary patterns, and it would be imprudent to draw from them broad conclusions about the relationship between music and language, but they do highlight an interesting type of question.

There are additional questions about the domain-specificity or domain-generality of grouping. On one view, the fact that grouping involves principles of audition independent of language means that it is not part of the narrow language faculty. While researchers are free to define technical terms as they see fit, it does seem to me that this notion of 'language faculty' is so narrow that it will fail to include the vast majority of all interesting cognitive processes involved in language, and may not include much of *anything* in the end. On the other hand, it is undoubtedly important, when asking about perceptual resources that music and language share, to bear in mind that those resources may also be shared with an array of perceptual processes in other domains and even other modalities. So there is good evidence that language and music are deeply similar with regard to grouping, but not that they are deeply different from other cognitive domains in this regard.

The presence of hierarchical structure is arguably less general and more difficult to explain than Gestalt grouping. One common view in language is that grouping 'inherits' its hierarchical structure from syntax, although there is significant disagreement on this point (and even less agreement about why *syntactic* structure is hierarchical). For music, Sect. 4.1 reviewed evidence that the relationship between harmonic syntax and rhythmic grouping is informationally 'non-directional'. That makes music consistent with the 'syntactic inheritance' view, though only if we accept musical syntax as hierarchically structured [see Katz (2017) and Temperley (2011) for arguments *pro* and *contra*, respectively]. On this view, the structural similarities between musical and linguistic grouping emerge from the fact that both involve translating between the hierarchical graph structure that represents the syntax of an utterance or piece of music and the temporal string of events that must be used to convey that structure from one organism to another. Hierarchically nested grouping should be shared with any cognitive domain that involves communication of hierarchically-structured representations through some sensory modality over time, including signed language and possibly dance in the visual modality. If some temporally complex cognitive activities lack hierarchical syntactic structure, there is no particular reason they should display hierarchical grouping. That said, it is quite difficult in practice to identify temporally complex cognitive activities that demonstrably lack hierarchical syntax. In principle, this view makes grouping relatively domain-specific, but with little evidence from external domains or even specifications of which domains *are* external.

Another possibility is that grouping hierarchy arises for reasons intrinsic to meter and rhythm. The study of temporal regularities at multiple timescales in language (e.g. Cummins and Port, 1998; Tilsen, 2009) and music (e.g. Jones and Boltz, 1989) has led to independent suggestions in the two domains that production and perception can be described with systems of hierarchically coupled oscillators. On this view, grouping is hierarchical because it is instantiated in individual brains, and brains are organized in terms of hierarchically coupled oscillators [see Hauk et al. (2017) for an overview oriented towards language]. As such, hierarchical grouping should be shared with all forms of motor control and temporally-modulated attending [see Tilsen (2009) for review of some relevant motor-control literature]. In principle, this view would make grouping relatively domain-general, but again, evidence from grouping structure in a broad array of cognitive domains

is not easy to find. Two relevant facts here are that prosodic groups clearly exist in languages without prosodic prominence, as detailed in Sect. 3; and that musical grouping exists in pieces without isochronous meter (see Popescu et al., 2021). So while hierarchical grouping may be related to hierarchical meter, it cannot be a direct consequence of meter.

In the end, then, what do we gain from the comparative study of computational-level musical and linguistic cognition? One answer is that simply attempting to align theories in the two domains helps clarify our thinking about each of them, especially at the 'architectural' level of information flow between components and the functional level of explaining why information in these domains is structured the way it is instead of some other way. A second answer is that, to the extent we can identify particular similarities and differences in the information states that underlie musical and linguistic cognition, those properties point to potentially fruitful areas of inquiry in other cognitive domains. And a final, optimistic answer is that any similarities may reveal deep cognitive properties rooted in evolution that distinguish human beings from other species. To make progress on any of these goals requires a sustained engagement with the analytical details and computational-level properties of a wide variety of cognitive resources underlying the composite concepts of 'language' and 'music'.

References

Agawu, V. K. (1990). Variation procedures in Northern Ewe song. *Ethnomusicology, 34*(2), 221–243.

Anderson, S. (1982). Where's morphology? *Linguistic Inquiry, 13*(4), 571–612.

Ayari, M., & McAdams, S. (2003). Aural analysis of Arabic improvised instrumental music (Taqsim). *Music Perception, 21*(2), 159–216.

Bagou, O., Fougeron, C., & Frauenfelder, U. (2002). Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. In *Speech Prosody 2002*, 159–162.

Baker, B. (2014). Word structure in Australian languages. In H. Koch & R. Nordlinger (Eds.), *The languages and linguistics of Australia: A comprehensive guide* (pp. 139–213). De Gruyter.

Beard, R. (1986). Neurological evidence for lexeme/morpheme-based morphology. *Acta Linguistica Academia Scientiarum Hungarica, 36*, 3–23.

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255–309.

Bhatara, A., Boll-Avetisyan, N., Agus, T., Höhle, B., & Nazzi, T. (2016). Language experience affects grouping of musical instrument sounds. *Cognitive Science, 40*, 1816–1830.

Blacking, J. (1970). Tonal organization in the music of two Venda initiation schools. *Ethnomusicology, 14*(1), 1–56.

Bouavichith, D., & Davidson, L. (2013). Segmental and prosodic effects on intervocalic voiced stop reduction in connected speech. *Phonetica, 70*, 182–206.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. The MIT Press.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics, 26*, 173–199.

Cambouropoulos, E. (2001). The local boundary detection model and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference* (pp. 17–22). ICMA.

Carlson, K., Clifton, C., & Frazier, L. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language, 45*, 58–81.

Charnavel, I. (2019). Steps towards a Universal Grammar of Dance: Local grouping structure in basic human movement perception. *Frontiers in Psychology, 10*, 1364. https://doi.org/10.3389/fpsyg.2019.01364.

Charnavel, I. (2022). Moving to the rhythm of spring: A case study of the rhythmic structure of dance. *Linguistics and Philosophy*. https://doi.org/10.1007/s10988-022-09356-z.

Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a, i/ in English. *Journal of the Acoustical Society of America, 117*(6), 3867–3878.

Choi, J. Y. (2003). Pause length and speech rate as durational cues for prosody markers. *Journal of the Acoustical Society of America, 114*(4), 2395.
Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics, 31*, 585–598.
Cohen Priva, U., & Gleason, E. (2020). The causal structure of lenition: A case for the causal precedence of durational shortening. *Language, 96*(2), 413–448.
Cole, R., & Cooper, W. (1975). Perception of voicing in English affricates and fricatives. *Journal of the Acoustical Society of America, 58*, 1280–1287.
Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics, 26*, 145–171.
De Jong, K. (2011). Flapping in American English. In M. Oostendorp, C. Ewen, & E. Hume (Eds.), *The Blackwell companion to phonology* (pp. 2711–2729). Wiley-Blackwell.
De Jong, K., & Zawaydeh, B. A. (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics, 27*, 3–22.
Deliège, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl and Jackendoff's grouping preference rules. *Music Perception, 4*, 325–360.
Dell, F. (2015). Text-to-tune alignment and lineation in traditional French songs. In T. Proto, P. Canettieri, & G. Valenti (Eds.), *Text and tune* (pp. 183–234). Peter Lang.
Dell, F., & Elmedlaoui, M. (2002). *Syllables in Tashlhiyt Berber and in Moroccan Arabic*. Kluwer.
Denham, S., & Winkler, I. (2014). Auditory perceptual organization. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of computational neuroscience*. Springer. https://doi.org/10.1007/978-1-4614-7320-6_100-1.
Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception and Psychophysics, 28*, 381–389.
Deutsch, D. (1999). Grouping mechanisms in music. In D. Deutsch (Ed.), *Psychology of music* (pp. 299–348). Academic Press.
DiCanio, C., Chen, W. R., Benn, J., Amith, J. D., & Castillo García, R. (2022). Extreme stop allophony in Mixtec spontaneous speech: Data, prosody, and modelling. *Journal of Phonetics, 92*, 101147.
Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics, 24*, 423–444.
Dowling, W. J. (1973). Rhythmic groups and subjective chunks in memory for melodies. *Perception and Psychophysics, 14*, 37–40.
Duanmu, S. (1996). Pre-juncture lengthening and foot binarity. *Studies in Linguistic Sciences, 26*, 95–115.
Ekwueme, L. E. N. (1975). Structural levels of rhythm and form in African music: With particular reference to the West Coast. *African Music, 5*(4), 27–35.
Ennever, T., Meakins, F., & Round, E. (2017). A replicable acoustic measure of lenition and the nature of variability in Gurindji stops. *Laboratory Phonology, 8*, 1–32.
Fenlon, J., & Brentari, D. (2021). Prosody: Theoretical and experimental perspectives. In J. Quer, R. Pfau, & A. Herrmann (Eds.), *Routledge handbook of theoretical and experimental sign language research.* Routledge. https://doi.org/10.4324/9781315754499-4.
Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review, 100*(2), 233–253.
Féry, C. (2010). Recursion in prosodic structure. *Phonological Studies, 13*, 51–60.
Féry, C., & Truckenbrodt, H. (2005). Sisterhood and tonal scaling. *Studia Linguistica, 59*, 223–243.
Fitch, T. (2015). The biology and evolution of musical rhythm: an update. In I. Toivonen, P. Csúri, & E. Van Der Zee (Eds.), *Structures in the mind: Essays on language, music, and cognition in honor of Ray Jackendoff* (pp. 293–324). MIT Press.
Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*, 3728–3740.
Frost, R., Monaghan, P., & Tatsumi, T. (2017). Domain-general mechanisms for speech segmentation: The role of duration information in language learning. *Journal of Experimental Psychology: Human Perception and Performance, 43*, 466–476.
Gallace, A., & Spence. C. (2011). To what extent do Gestalt grouping principles influence tactile perception? *Psychological Bulletin, 137*(4), 538–561.
Gee, J. P., & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology, 15*, 411–458.
Gordon, M., & Munro, P. (2007). A phonetic study of final vowel lengthening in Chickasaw. *International Journal of American Linguistics, 73*, 293–330.
Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America, 47*, 613–617.

Halle, J. (2004). *Constituency matching in metrical texts*. Yale University.
Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *The view from Building 20* (pp. 111–176). MIT Press.
Harwood, D. L. (1976). Universals in music: A perspective from cognitive psychology. *Ethnomusicology, 20*(3), 521–533.
Hauk, O., Giraud, A., & Clarke, A. (2017). Brain oscillations in language comprehension. *Language, Cognition and Neuroscience, 32*, 533–535.
Hayes, B. (1989). The prosodic hierarchy in meter. *Phonetics and Phonology, 1*, 201–260.
Hayes, B., & Lahiri, A. (1991). Bengali intonational phonology. *Natural Language and Linguistic Theory, 9*, 47–96.
Hayes, B., & Schuh, R. G. (2019). Metrical structure and sung rhythm of the Hausa rajaz. *Language, 95*(2), 253–299.
Heffner, C., & Slevc, R. (2015). Prosodic structure as a parallel to musical structure. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01962.
Higgins, K. M. (2006). The cognitive and appreciative import of musical universals. *Revue Internationale de Philosophie, 4*, 487–503.
Hockett, C. (1960). The origin of speech. *Scientific American, 203*, 88–111.
Horn, E. (2010). *Poetic organization and poetic license in the lyrics of Hank Williams, Sr. and Snoop Dogg*. PhD thesis, University of Texas at Austin.
Hualde, J. I., Simonet, M., & Nadeu, M. (2011). Consonant lenition and phonological recategorization. *Laboratory Phonology, 2*, 301–329.
Hughes, D. W. (1988). Deep structure and surface structure in Javanese music: A grammar of Genghing Lampah. *Ethnomusicology, 32*(1), 23–74.
Huron, D. (1991). Tonal consonance versus tonal fusion in polyphonic sonorities. *Music Perception, 9*, 135–154.
Hyman, L. (2013). Penultimate lengthening in Bantu. In B. Bickel, L. Grenoble, D. Peterson, & A. Timberlake (Eds.), *Language typology and historical contingency* (pp. 309–330). John Benjamins.
Ishihara S. (2003). *Intonation and interface conditions*. PhD thesis, MIT.
Iverson, J., Patel, A., & Ohgushi, K. (2008). Perception of rhythmic grouping depends on auditory experience. *Journal of the Acoustical Society of America, 124*, 2263–2271.
Jackendoff, R. (2002). *Foundations of language*. Oxford University Press.
Johnson, K., & Martin, J. (2001). Acoustic vowel reduction in Creek: Effects of distinctive length and position in the word. *Phonetica, 58*, 81–102.
Johnson-Laird, P. (1991). Jazz improvisation: A theory at the computational level. In P. Howell, R. West, & I. Cross (Eds.), *Representing musical structure* (pp. 291–326). Academic Press.
Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review, 96*, 459–491.
Jun, S. (1993). *The phonetics and phonology of Korean prosody*. PhD thesis, Ohio State University.
Jusczyk, P., & Krumhansl, C. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 627–640.
Kaliakatsos-Papakostas, M., Katsiavalos, A., Tsougras, C., & Cambouropoulos, E. (2014). Harmony in the polyphonic songs of Epirus: Representation, statistical analysis and generation. In A. Holzapfel (Ed.), *Proceedings of the 4th International Workshop on Folk Music Analysis* (pp. 21–28). Bogazici University.
Katz, J., & Pesetsky, D. (2011). The identity thesis for language and music. *LingBuzz*: lingbuzz/000959.
Katz, J. (2015). Hip-hop rhymes reiterate phonological typology. *Lingua, 160*, 54–73.
Katz, J. (2016). Lenition, perception, and neutralisation. *Phonology, 33*, 43–85.
Katz, J. (2017). Harmonic syntax of the 12-bar blues: A corpus study. *Music Perception, 35*, 165–192.
Katz, J. (2021). Intervocalic lenition is not phonological: Evidence from Campidanese Sardinian. *Phonology, 38*(4), 651–692.
Katz, J., & Fricke, M. (2018). Auditory disruption improves word segmentation: a functional basis for lenition phenomena. *Glossa, 3*(1), 38.
Katz, J., & Pitzanti, G. (2019). The phonetics and phonology of lenition: A Campidanese Sardinian case study. *Laboratory Phonology, 10*(1), 1–40.
Keating, P., Cho, T., Fougeron, C., & Hsu, C. (2003). Domain-initial strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology VI* (pp. 143–161). Cambridge University Press.
Kim, S. (2004). *The role of prosodic phrasing in Korean word segmentation*. PhD thesis, University of California.

Kingston, J. (2008). Lenition. In L. Colantoni, & J. Steele (Eds.), *Proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology* (pp. 1–31). Cascadilla Proceedings Project.
Kirchner, R. (1998). *An effort-based approach to consonant lenition*. PhD thesis, University of California.
Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics, 35*, 162.
Krumhansl, C., & Jusczyk, P. (1990). Infants' perception of phrase structure in music. *Psychological Science, 1*, 70–73.
Ladd, D. R. (1986). Intonational phrasing: The case for recursive prosodic structure. *Phonology Yearbook, 3*, 311–340.
Ladd, D. R. (2012). What *is* duality of patterning, anyway? *Language and Cognition, 4*, 261–273.
Ladefoged, K., & Johnson, K. (2011). *A course in phonetics*. Wadsworth.
Langus, A., Seyed-Allaei, S., Uysal, E., Pirmoradian, S., Marino, C., & Nespor, M. (2016). Listening natively across perceptual domains? *Journal of Experimental Psychology: Learning, Memory and Cognition, 42*, 1127–1139.
Large, E. W., Palmer, C., & Pollack, J. B. (1995). Reduced memory representations for music. *Cognitive Science, 19*, 53–96.
Lavoie, L. (2001). *Consonant strength: Phonological patterns and phonetic manifestations*. Garland.
Lerdahl, F. (2001). *Tonal pitch space*. Oxford University Press.
Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. MIT Press.
Marr, D. (1982). *Vision*. MIT Press.
Mattys, S. L., & Jusczyk, P. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*, 91–121.
McCawley, J. (1968). *The phonological component of a grammar of Japanese*. Mouton.
McPherson, L., & Ryan, K. M. (2018). Tone-tune association in Tomma So (Dogon) folk songs. *Language, 94*(1), 119–156.
McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39*, 21–46.
Meyer, L. B. (1991). A pride of prejudices: Or, delight in diversity. *Music Theory Spectrum, 13*(2), 241–251.
Millotte, S., Morgan, J., Margules, S., Bernal, S., Dutat, M., & Christophe, A. (2011). Phrasal prosody constrains word segmentation in French 16-month-olds. *Journal of Portuguese Linguistics, 9*, 67–86.
Molnar, M., Carreiras, M., & Gervain, J. (2016). Language dominance shapes non-linguistic rhythmic grouping in bilinguals. *Cognition, 152*, 150–159.
Morén, B., & Zsiga, E. (2006). The lexical and post-lexical phonology of Thai tones. *Natural Language & Linguistic Theory, 24*, 113–178.
Mungan, E., Yazici, Z. F., & Kaya, M. (2017). Perceiving boundaries in unfamiliar Turkish Makam music: Evidence for Gestalt universals? *Music Perception, 34*(3), 267–290.
Nagano-Madsen, Y. (1992). *Mora and prosodic coordination: A phonetic study of Japanese, Eskimo, and Yoruba*. Lund University Press.
Nakatani, L., & Dukes, K. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America, 62*, 714–719.
Nan, Y., Knösche, T. R., & Friederici, A. D. (2009). Non-musicians' perception of phrase boundaries in music: A cross-cultural ERP study. *Biological Psychology, 82*, 70–81.
Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realisation model*. University of Chicago Press.
Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Foris.
Nordhoff, S. (2012). Synchronic grammar of Sri Lanka Malay. In S. Nordhoff (Ed.), *The genesis of Sri Lanka Malay: A case of extreme language contact* (pp. 13–52). Brill.
Oller, D. (1979). Syllable timing in Spanish, English, and Finnish. In P. Macneilage (Ed.), *Current issues in the phonetic sciences* (pp. 189–216). Springer.
Onaka, A., Palethorpe, S., Watson, C., & Harrington, J. (2003). Acoustic and articulatory difference of speech segments at different prosodic positions. In C. Bow (Ed.), *Proceedings of the 9th Australian International Conference on Speech Science and Technology* (pp. 148–153). ASSTA.
Ortega-Llebaria, M., & Prieto, P. (2010). Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Language and Speech, 54*, 73–97.
Palmer, C., & Krumhansl, C. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 728–741.
Pasciak, K. J. (2017). *A transformational approach to Japanese traditional music of the Edo period*. PhD thesis, UMass Amherst.

Patel-Grosz, P., Grosz, P., Kelkar, T., & Jensenius, A. (2018). Coreference and disjoint reference in the semantics of narrative dance. *Proceedings of Sinn und Bedeutung*, *22*(2), 199–216. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/78.

Pearce, M. (2008). The perception of grouping boundaries in music. Queen Mary University (Unpublished manuscript).

Peretz, I. (1989). Clustering in music: An appraisal of task factors. *International Journal of Psychology, 24*, 157–178.

Pierrehumbert J. (1980). *The phonology and phonetics of English intonation*. PhD thesis, MIT.

Pierrehumbert, J., & Beckman, M. (1988). *Japanese tone structure*. MIT Press.

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

Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. Docherty & D. R. Ladd (Eds.), *Papers in laboratory phonology II* (pp. 90–117). Cambridge University Press.

Popescu, T., Widdess, R., & Rohrmeier, M. (2021). Western listeners detect boundary hierarchy in Indian music: A segmentation study. *Scientific Reports, 11*, 3112.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90*, 2956–2970.

Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics, 29*, 407–429.

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's 'Träumerei.' *Journal of the Acoustical Society of America, 92*, 2546–2568.

Repp, B. H. (1998). A microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America, 104*, 1085–1100.

Richards, N. (2010). *Uttering trees*. MIT Press.

Richards, N. (2016). *Contiguity theory*. MIT Press.

Rohrmeier, M. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music, 5*, 35–53.

Rohrmeier, M., Zuidema, W., Wiggins, G., & Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society B, 370*, 20140097.

Rohrmeier, M. (2020a). The syntax of jazz harmony: Diatonic tonality, phrase structure, and form. *Music Theory & Analysis, 7*(1), 1–62.

Rohrmeier, M. (2020b). Towards a formalization of musical rhythm. In J. Cumming et al. (Eds.), *Proceedings of the 21st International Society for Music Information Retrieval Conference* (pp. 621–629). ISMIR.

Saffran, J., Aslin, R., & Newport, E. (1996a). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Saffran, J., Newport, E., & Aslin, R. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory & Language, 35*, 606–621.

Schafer, A., Speer, S., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research, 29*, 169–182.

Schellenberg, E. G. (1996). Expectancy in melody: Tests of the implication-realization model. *Cognition, 58*, 75–125.

Schlenker, P. (2017). Outline of music semantics. *Music Perception, 35*(1), 3–37.

Selkirk, E. (1972). *The phrase phonology of English and French*. Garland.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. MIT Press.

Selkirk, E. (2011). The syntax-phonology interface. In J. Goldsmith, J. Riggle, & A. Yu (Eds.), *The handbook of phonological theory* (pp. 435–483). Blackwell.

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*, 193–247.

Shaw, J., Carignan, C., Agostini, T., Mailhammer, R., Harvey, M., & Derrick, D. (2020). Phonological contrast and phonetic variation: The case of velars in Iwaidja. *Language, 96*(3), 578–617.

Spelke, E. (1994). Initial knowledge: Six suggestions. *Cognition, 50*, 431–445.

Steedman, M. (1984). A generative grammar for jazz chord sequences. *Music Perception, 2*, 52–77.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry, 31*, 649–689.

Stevens, C., & Byron, T. (2009). Universals in music processing. In S. Hallam, I. Cross, & M. Thaut (Eds.), *Oxford handbook of music psychology* (pp. 19–31). Oxford University Press.

Stock, J. (1993). The application of Schenkerian analysis to ethnomusicology: Problems and possibilities. *Music Analysis, 12*(2), 215–240.
Stoffer, T. H. (1985). Representation of phrase structure in the perception of music. *Music Perception, 3*, 191–220.
Temperley, D. (2000). Meter and grouping in African music: A view from music theory. *Ethnomusicology, 44*(1), 65–96.
Temperley, D. (2011). Composition, perception, and Schenkerian theory. *Music Theory Spectrum, 33*, 146–168.
Thom, B., Spevak, C., & Höthker, K. (2002). Melodic segmentation: Evaluating the performance of algorithms and musical experts. In *Proceedings of the 2002 International Computer Music Conference* (pp. 65–72). ICMA.
Thorpe, L. A., Trehub, S. E., Morrongiello, B. A., & Bull, D. (1988). Perceptual grouping by infants and preschool children. *Developmental Psychology, 24*(4), 484–491.
Tillmann, B., & McAdams, S. (2004). Implicit learning of musical timbre sequences: Statistical regularities confronted with acoustic (dis)similarities. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*, 1131–1142.
Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science, 33*, 839–879.
Tingsabadh, M. R. K., & Abramson, A. (1993). Thai. *Journal of the International Phonetic Association, 23*, 24–28.
Todd, N. (1985). A model of expressive timing in tonal music. *Music Perception, 3*, 33–57.
Trehub, S. (2000). Human processing predispositions and musical universals. In N. Wallin, B. Merker, & S. Brown (Eds.), *The Origins of music* (pp. 427–448). MIT Press.
Turk, A., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics, 28*, 397–440.
Turpin, M. (2007). The poetics of Central Australian song. *Australian Aboriginal Studies, 2*(2), 100–115.
Turpin, M. (2011). Artfully hidden: Text and rhythm in a central Australian aboriginal song series. *Musicology Australia, 29*(1), 93–108.
Turpin, M. (2017). Parallelism in Arandic song-poetry. *Oral Tradition, 31*(2), 535–560.
Van der Hulst, H., & Gordon, M. (2020). Word stress systems. In C. Gussenhoven & A. Chen (Eds.), *The Oxford handbook of language prosody* (pp. 66–77). Oxford University Press.
Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin, 138*(6), 1172–1217.
Wagner, M. (2005). *Prosody and recursion*. PhD thesis, MIT.
Wagner, M. (2022). Two-dimensional parsing of the acoustic stream explains the Iambic-Trochaic Law. *Psychological Review*, *129*(2), 268–288. https://doi.org/10.1037/rev0000302.
Wagner, M., & McAuliffe, M. (2019). The effect of focus prominence on phrasing. *Journal of Phonetics, 77*, 100930.
Wertheimer, M. (1938). Laws of organization in perceptual forms [English translation of 1923 essay]. In W. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 71–88). Routledge.
Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*, 1707–1717.