Interpreting phonetic evidence for hierarchical organization of prosodic phrases

Seung-Eun Kim*, Sam Tilsen

Cornell University
203 Morrill Hall, Ithaca, NY 14853
*Corresponding author: sk2996@cornell.edu

Abstract

A fundamental issue in models of speech production is how differences in syntactic structure are manifested as prosodic variation. Many prosodic theories presuppose that syntactic differences are mapped to hierarchically organized prosodic phrase structure. These theories predict that acoustic and articulatory measurements, particularly at phrase boundaries, should reflect categorical differences in abstract levels of phrasal organization. At the same time, it is widely accepted that factors such as speech rate may modulate the syntax-prosody mapping. We argue that the existing phonetic evidence for hierarchical organization of prosodic phrases is ambiguous, and that a non-hierarchical organization of phrases is also consistent with the evidence. To compare hierarchical and non-hierarchical organization models, the current study elicited productions of English non-restrictive and restrictive relative clauses (NRRC vs. RRC) at varying speech rates. Articulatory and acoustic variables associated with prosodic boundaries were analyzed, incorporating speech rate as a covariate. Overall, the evidence for multiple levels of prosodic phrase categories was not very compelling. The measures that were most supportive of hierarchical phrase structure were measures of boundary-related slowing and gestural overlap at boundaries. Notably, patterns of evidence differed across participants and also according to the scale on which speech rate is defined.

Keywords: hierarchical prosodic structure, prosodic phrases, phonetic evidence, syntax-prosody mapping, speech rate, mixture regression models

1. Introduction

It is often taken for granted that different syntactic structures are associated with categorically distinct prosodic organizations. In particular, it is argued that syntactic differences can correspond to differences in hierarchical phrase structure, which in turn may be associated with categorical variation in boundary strengths or intonational patterns. A plausible alternative, however, is that prosodic words are grouped into non-hierarchically organized phrases, and variation in phonetic measures at phrase boundaries is gradient and conditioned on a variety of factors. The aim of this paper is to step back from the common assumption that there necessarily exist a hierarchy of prosodic phrase types and to critically assess whether phonetic observations can be used to draw strong inferences about the nature of prosodic phrase structure.

To accomplish this, analyses were conducted on articulatory and acoustic data collected from an experiment in which two types of English relative clauses with nearly identical lexical content were elicited. These were restrictive relative clauses (RRC), e.g. *The Mr. Hodd who knows Mr. Robb plays tennis*, and non-restrictive relative clauses (NRRC), e.g. *A Mr. Hodd, who knows Mr. Robb, plays tennis* (see Arnold, 2007 for semantic and syntactic analyses). A novel method for eliciting continuous variation in speech rate was employed in order to obtain productions of these sentences at varying rates. We analyzed whether articulatory and acoustic variables differ by syntactic context and whether such differences interact with speech rate. We also examined whether phonetic variables exhibit evidence of speech rate-dependent mixtures of categories within a given syntactic context. In each of these analyses, we compared the use of the two different types of global speech rate measures – inverse rate (sentence duration) and proper rate (the reciprocal of sentence duration).

The main finding was that the experimental data do not provide unambiguous evidence for a hierarchy of phrasal categories; such evidence was inconsistent across participants, contexts, and phonetic variables. The phonetic variables that most frequently showed evidence for categorically distinct phrase types were those related to articulatory slowing and gestural overlap at phrases boundaries. Interestingly, the choice of rate measure had some effects on observed relations between speech rate and dependent variables.

The analyses presented here are exploratory in the sense of Baayen et al. (2017); namely, we are primarily concerned with model criticism and interpretation, as opposed to confirmatory hypothesis testing. As advocated in a recent issue of this Journal (Roettger et al., 2019), exploratory analyses are a necessary precursor to confirmatory analyses and should not be undervalued. Some additional contributions of this paper are a generic analysis of how phonetic measures can be used as evidence for hierarchical prosodic structure and an examination of how the choice of rate measure can influence analyses.

The paper is structured as follows. Section 2.1 describes the syntactic/semantic contexts employed in the experiment, hierarchical and non-hierarchical prosodic organizations and associated phonetic measures, and the nature of syntax-prosody mapping. Section 2.2 examines different types of phonetic evidence for the existence of hierarchical prosodic organizations and proposes a range of the quality of such evidence. Section 2.3 discusses why it is important to consider how quantitative measures of rate are expressed. Section 2.4 presents the hypotheses and delineates the exploratory aspects of our analyses. Section 3 details the experimental design and analyses methods. Section 4 presents the results, and Section 5 provides further discussion.

2. Background

2.1. Syntactically conditioned variation and the syntax-prosody mapping

To elicit prosodic variation, the current study makes use of the syntactic contrast between restrictive and non-restrictive relative clauses in English. Examples of these are shown in Table 1. A non-restrictive relative clause (NRRC) contributes information regarding the referent of the expression that it modifies (here a person, Mr. Hodd), but the information is not essential to identifying that referent. NRRCs are often separated in orthography from a main clause by commas, dashes, or parentheses. Native speakers have the intuition that an NRRC can be produced as an "aside" or parenthetical, and silent pauses are felicitous before and after the relative clause. In contrast, the information provided by a restrictive relative clause (RRC) is essential to identifying a referent from a set of possible referents: for example, there could be two different people who are named Mr. Hodd, and the RRC picks out one in particular. It is less natural to pause before or after RRCs, and it is unusual to separate them with commas or other punctuation in orthography.

Table 1. Examples of non-restrictive and restrictive relative clauses and accompanying contexts used in the experiment.

Context:	There is one Mr. Hodd. He knows Mr. Robb.
Non-restrictive relative clause (NRRC):	A Mr. Hodd, who knows Mr. Robb, often plays tennis.
Context:	There are two Mr. Hodds. Only one knows Mr. Robb.
Restrictive relative clause (RRC):	The Mr. Hodd who knows Mr. Robb often plays tennis.

There are a number of syntactic tests which have been argued to distinguish NRRCs and RRCs (e.g. Demirdache, 1991; Fabb, 1990; McCawley, 1981) and there is some agreement on their semantic and pragmatic differences (e.g. Del Gobbo, 2007). However, there does not exist a consensus on how the difference should be represented syntactically (see Arnold, 2007 for an overview). Nonetheless, it is presupposed here that the contrast between RRC and NRRC is a categorical difference in syntactic structure, and we are interested in how to conceptualize the prosodic manifestation of this difference.

There are a variety of ways to conceptualize prosodic organization. One of the most well-known analyses comes from Selkirk (2005), who uses hierarchical phrase structure representations of the sort in Fig. 1. For example, in Fig. 1A, the main clause subject and relative clause are separate intermediate phrases (ip), and together constitute an entire intonational phrase (IP). In Fig. 1B, the main clause subject and relative clause together comprise a single intermediate phrase (ip) and are part of an intonational phrase along with the main clause predicate. We refer to this sort of analysis as hierarchical because it posits more than one "level" or "type" of phrase (i.e. ip, IP). Note that the issue of whether there is a hierarchical organization of units below the phrase (i.e. prosodic words, feet, syllables) is orthogonal to our concerns here.

In any phrase structure, it is possible to identify entities which we refer to as *boundaries* (or alternatively, *domain edges*). These boundaries can be associated with the brackets in the text below the phrase structures in Fig. 1. It is important to note that a "boundary" is a different sort of entity than the prosodic units which constitute the hierarchy; it is not sensible to posit phrasal boundaries in the absence of some sort of phrasal organization, although that organization need not be hierarchical.

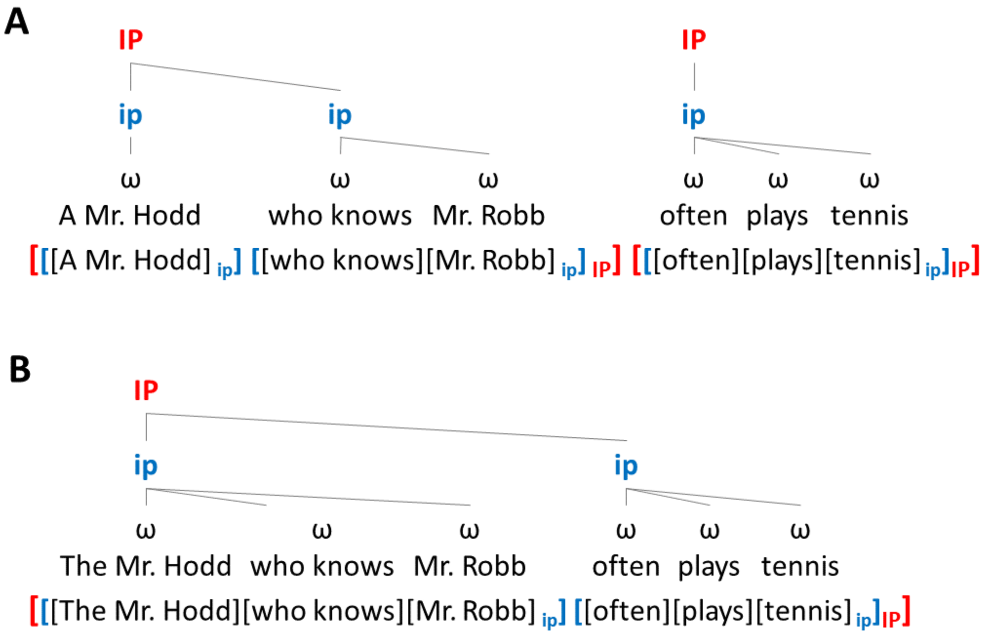
A
IP
ip
ip
ω
ω
ω
A Mr. Hodd
who knows
Mr. Robb
IP
ip
ω
ω
ω
often
plays
tennis
[[[A Mr. Hodd] ip] [[who knows][Mr. Robb] ip] IP] [[[often][plays][tennis] ip]IP]
B
IP
ip
ip
ω
ω
ω
The Mr. Hodd
who knows
Mr. Robb
ω
ω
ω
often
plays
tennis
[[[The Mr. Hodd][who knows][Mr. Robb] ip] [[often][plays][tennis] ip]IP]

Fig. 1. Examples of hierarchical organizations of prosodic phrases, following Selkirk (2005). (A) The main clause subject and relative clause are separate intermediate phrases (ip), and the subject and verb phrases are separate intonational phrases (IP). (B) The main clause subject and relative clause are a single ip, and the whole sentence is a single IP.

Questions regarding the ontological status of boundaries are often sidestepped in recent experimental studies, but early theoretical literature took seriously the possibility that representations include “boundary symbols” (cf. Chomsky & Halle, 1968; McCawley, 1968; Selkirk, 1972). Boundary symbols in such approaches are functionally equivalent to segments in that they can comprise the environments of phonological rules. Selkirk (1980) and later Hayes (1989) made various arguments that boundary symbols are overly powerful and are unnecessary if a hierarchical prosodic structure is assumed. A number of theories impose or relax various constraints on representations of hierarchical prosodic structure. For example, some argue that the levels are strictly layered, such that a given phrasal category always dominates only categories on the next lowest level (Nespor & Vogel, 1986; Selkirk, 1984). A related issue is whether, if not strictly layered, phrasal categories can recursively dominate instances of themselves (e.g. Ladd, 1986; Wagner, 2005). Moreover, a variety of theorists have argued for a greater number of phrase categories, or language-specific categories (see Jun, 2005). In most cases, the arguments for or against various conceptualizations are based on phonological patterns.

The aim of this paper is not to resolve among these various possibilities. Furthermore, our analysis below does not address phonological arguments, such as observations that phonological alternations may be conditioned on hypothesized prosodic contexts. Using phonological patterns to determine prosodic structure may be problematic if the phonological patterns are probabilistic. To our knowledge, there is no way to obtain a ground truth on how the words in an utterance are organized into phrases. Thus, arguments for the existence of hierarchical phrase structures which are based on correlations between phonological patterns and those hypothesized structures are circular: the existence of the hierarchical organization is presupposed in the argument itself.

Instead, we focus on interpreting phonetic measures as evidence for hierarchically organized prosodic phrases. An important point to make at the outset is that there are two unknown links between syntactic structure and phonetic measures. First, there is the syntax-prosody mapping, whereby syntactic structure may be associated with a prosodic organization. Second, there is the prosody-articulation mapping, whereby prosodic organizations of the sort in Fig. 1 (or parameters associated prosodic organizations in Fig. 2) influence the control of articulation. Formal representations of prosodic organization do not provide explicit models of how those representations are translated to articulatory processes in speech, and thus a number of assumptions are required to interpret the phonetic measures that might differentiate them. Yet, it stands to reason that if there does exist a hierarchical organization of phrase types, there is a potential for those differences to be detectable in distributions of phonetic measurements.

Based on previous studies, there are a variety of acoustic and articulatory measures which may correlate with differences in prosodic organization. Variation in such measures is commonly held to arise from categorical differences in hierarchical prosodic phrase structure. Yet, for reasons that we discuss below, this interpretation is not wholly justified.

One robust correlate of prosodic variation is phrase-final lengthening, which refers to the temporal extension of acoustic and articulatory intervals near the ends of phrases. Phrase-final lengthening has been found to correlate with hypothesized variation in the "strength" of the boundary (e.g. Byrd & Saltzman, 1998; Price et al., 1991; Wightman et al., 1992). Note that it is not necessary to adopt a categorical interpretation of "strength" in order to observe such effects. The durational differences associated with phrase-final lengthening may be observed most strongly at the rime of the phrase-final syllable (e.g. Turk & Shattuck-Hufnagel, 2007; Wightman et al., 1992), and possibly, more robust differences will be found at segments closer to a boundary (Berkovits, 1993a, 1993b). Relatedly, the likelihood of a pause and pause duration are associated with prosodic variation. In hierarchical frameworks, it has been argued that pauses may follow IP boundaries, but not lower-level boundaries such as ip (Nespor & Vogel, 1986). However, there are corpus studies which contradict this argument; such studies show that pause likelihood and duration vary with boundary strength (Choi, 2003; Horne et al., 1995), but do not necessitate a hierarchical phrase structure model.

Second, differences in prosodic organization also manifest in articulatory measures. The timing intervals between articulatory gestures have been found to vary in relation to hypothesized differences in prosodic organization, and these effects have been observed both at the right and left edges of the prosodic domains (e.g. Byrd, 2000; Byrd & Saltzman, 1998; Cho, 2006; Cho & Keating, 2001; Keating et al., 2004). As in acoustic durations, movement durations will differ more for gestures closer to a boundary (Byrd et al., 2006). Furthermore, articulatory targets may reflect prosodic organization, as higher level domains have been associated with more extreme articulatory targets than lower domains (e.g. Cho & Keating, 2001; Fougeron & Keating, 1997; Keating et al., 2004).

Another possible articulatory correlate of prosodic organization is peak movement velocity. However, studies have shown conflicting results (e.g. Cho, 2006; Edwards et al., 1991; Tabain, 2003). Specifically, Edwards et al. (1991) found slower peak velocity as the strength of prosodic boundary increases, while Tabain (2003) observed contrary results such that the peak velocity increased at higher prosodic domains. Cho (2006) examined kinematic variations in a $C_1V_1\#C_2V_2$ sequence where C1 and C2 were always /b/ and V1 and V2 were either /i/ or /a/; prosodic boundary (#) varied in three levels (prosodic word, ip, IP). The peak velocity did not vary according to prosodic boundary strength in phrase-final (C_1V_1) or phrase-initial (C_2V_2) lip opening movements, but it decreased at higher prosodic domains in transboundary lip closing movements ($V_1\#C_2$).

Lastly, fundamental frequency (F0), a correlate of intonation, also reflects variation in prosodic organization. A number of studies have observed correlations between phrasal organizations and pitch accent scaling or pitch reset (e.g. Fery & Truckenbrodt, 2005; Ladd, 1988; Truckenbrodt, 2002; van den Berg et al., 1992). For instance, Ladd (1988) examined the intonation patterns of contrasting English coordinative structures, i.e. (1) A but [B and C] vs. (2) [A and B] but C. He found that in structure (1), C had a lower F0 than B, and B had a lower F0 than A, whereas in structure (2), B had a lower F0 than A, but C was lowered relative to A not B. Thus, F0 peaks in C were higher in structure (2) than in structure (1). This shows that differences in phrasal organization are reflected intonationally.

Although many previous studies have presupposed hierarchical prosodic organizations such as in Fig. 1, we do not take hierarchical phrase structure for granted. An alternative possibility shown in Fig. 2 is that differences in syntactic organization and other factors are manifested as gradient variation of parameters associated with a non-hierarchical organization of phrases. This non-hierarchical organization could be common to different syntactic constructions. For example, NRRCs and RRCs may be produced with identical, non-hierarchically organized phrase structures, as shown Fig. 2A and B. The mechanisms which control this phrasal organization may allow for local, gradient modulations of articulatory processes, particularly at phrase boundaries. No categorical distinction in hierarchical prosodic structure exists between Fig. 2A and B, and yet various factors—including but not limited to syntactic structure—might influence local modulations of articulation and intonation at boundaries such that phonetic measures correlate with the two syntactic constructions. Note that this conception assumes a phrase-building or chunking mechanism which operates on prosodic words but does not prescribe any further hierarchical organization of phrases. It allows for the order of chunks (i.e. phrases) to have an influence on phonetic measures but does not posit a hierarchy of phrase types. We do not rule out this possibility a priori, even though it appears to have been implicitly rejected by many studies of the syntax-prosody interface.

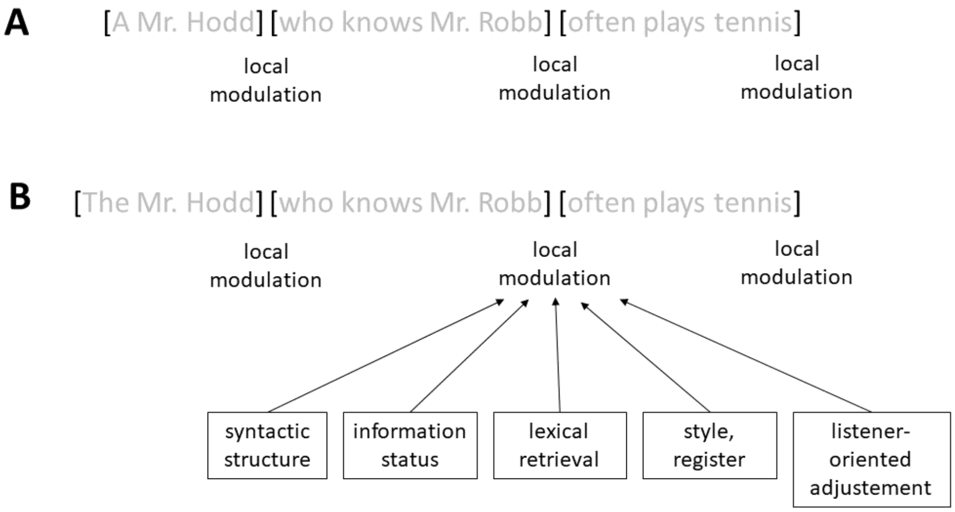

Fig. 2. An example of non-hierarchical prosodic organization, where utterances are chunked into phrases. Local modulation of articulatory processes at the ends of phrases can be influenced by various factors.

Many of the studies cited above have taken the existence of a hierarchy of phrasal categories or domains for granted. For another example, consider the study of informational effects on prosody in Aylett and Turk (2004). The Aylett and Turk study conducts regression analyses of syllable durations at boundaries, and the authors impose a hierarchy of prosodic units that include prosodic words, minor phrases, major phrases, and full intonational phrases. Not surprisingly, the analysis finds an effect of hierarchically organized prosodic boundary (treated as a categorical variable). Should this finding be taken as evidence about the ontological status of boundaries, such that there are indeed several categorically distinct prosodic boundaries? We do not think so, for the following reasons.

First, note that the boundaries were identified via a ToBI-based hand-coding procedure, which is known to have only moderate intercoder reliability, especially on markings of the *type* or *identity* of intonation labels (e.g. Syrdal & McGory, 2000; Wightman, 2002). If there were categorically distinct types of phrases, one might expect speakers to be aware of and exhibit a high level of agreement upon those categories, in the same way that speakers have a high level of agreement upon the phonemic categories in word forms.

Second, consider that the same regression effects could arise in the absence of categorically different phrase types if there exists a single "type" of phrase whose phonetic effects are gradiently parameterized, as suggested in Fig. 2. The gradient variation might be modulated by a variety of factors that include syntactic structure, pragmatic/informational context, and local speech rate modulations. Such factors are likely to have a strong influence on the hand coding procedure. Indeed, in the same study, Aylett and Turk (2004) found that informational factors such as log word frequency, syllable trigram probability, and givenness also have effects on syllable duration at phrase boundaries. From an analytical perspective, if informational factors are correlated with the hand-coded boundary strength labels, then any regression with both sets of factors will not accurately estimate coefficients due to collinearity (Tomaschek et al., 2018), and separate regressions with prosodic or informational predictors cannot resolve independent contributions of these factors. Hence the data cannot be used to arrive at the conclusion that there are categorically different levels of phrases with categorically different types of boundaries. Notably, the Aylett and Turk study is not unique in appearing to take the ontological status of prosodic phrasal categories for granted.

To minimize the number of pre-suppositions in our interpretation, we strive to be explicit regarding assumptions about the nature of the syntax-prosody mapping. To that end, it is necessary to distinguish between deterministic and probabilistic views of the mapping. In a probabilistic mapping, each member of the set of prosodic phrase categories is assigned a conditional probability. For example, if there are two possible prosodic structures X_1 and X_2, a given syntactic category A would be implemented as prosodic structure X_1 with probability $\mathrm{p}(X_1|A)$ and as structure X_2 with probability $p(X_2|A) = 1 - p(X_1|A)$.

It is also important to distinguish between categorical and gradient forms of prosodic variation. A hierarchical model of phrase organization entails that there should be categorical variation in the strengths of phrase boundaries, which corresponds to hypothesized phrase categories such as ip or IP in Fig. 1. In contrast, a non-hierarchical model of phrasal organization entails that boundary strengths differ only gradiently between syntactic contexts—there are not different "types" of phrase categories in such a view; rather one could imagine a numeric parameter (or set of such parameters) whose values are determined by a variety of factors, as in Fig. 2. Note that our use of "organization" does not entail categorical differences but can equally refer to gradient prosodic variations.

The combination of probabilistic vs. deterministic distinction and categorical vs. gradient distinction leads to the four logically possible forms of syntax-prosody mapping listed in Table 2, arranged according to the form of prosodic variation (categorical vs. gradient) and the determinacy of the syntax-prosody mapping (probabilistic vs. deterministic). Table 2(a) and (b) both represent mappings in which variation in syntactic structure is correlated with categorical variation in prosodic phrase structure; they differ in terms of whether syntactic contexts are deterministically or probabilistically mapped to prosodic structures. In the gradient mappings of (c) and (d), different syntactic contexts map to gradient parameter values which would be associated with the non-hierarchical organization shown in Fig. 2. For the probabilistic mapping (c), the form of the mapping could be described by conditional probability distributions over parameter values for each syntactic context. For the deterministic mapping (d), different syntactic contexts would be associated with specific parameter values.

Table 2. Logically possible forms of syntax-prosody mapping

	determinacy of syntax-prosody mapping	
form of prosodic variation	(a) categorical, probabilistic	(b) categorical, deterministic
	(c) gradient, probabilistic	(d) gradient, deterministic

The range of logical possibilities in Table 2 suggests that in interpreting phonetic measures associated with different syntactic constructions (e.g. RRCs vs. NRRCs), one should not assume that observations associated with a given syntactic construction are always associated with the same prosodic organization. Factors such as speech rate, inter-participant variation, practice effects, and fluctuations in participant arousal may influence the probability of a given prosodic parameter or category, even within the same syntactic context.

2.2. Assessing phonetic evidence for hierarchical prosodic phrase structure

How can we tell if values of a phonetic variable provide evidence for a hierarchical organization of prosodic phrases? Here we argue that there is no way to find unambiguous evidence for such organization in phonetic measurements. However, we propose that there is a range of the quality of evidence that can be obtained, with some forms of evidence being relatively weak, and other forms being more compelling. Analyses which take speech rate into account can provide better evidence than analyses that do not, and analyses which do *not* rely on a syntactic or semantic contrast are more powerful than those that do. We will discuss the relative quality of three different types of evidence in this section.

The first type of evidence is a syntactically-conditioned difference in the distributions of phonetic variables. Although we examine whether such differences exist in our data, we argue that this evidence does not provide strong support for the existence of distinct prosodic categories. To illustrate this point, Fig. 3 (i) and (ii) compare hypothesized distributions of a phonetic variable observed in arbitrary syntactic constructions A and B. We refer to comparisons of this sort as *paradigmatic comparisons*, because they involve a comparison of variables associated with similar word sequences in different syntactic constructions. A correlation between syntactic contexts and values of a phonetic variable tells us nothing specific about the organization of phrases or the mechanisms which are responsible for the correlation. It could be the case that the syntactic constructions map to distinct prosodic organizations, $\Phi 1$ and $\Phi 2$, associated with different phonetic parameter values, $\mu_{\Phi 1}$ and $\mu_{\Phi 2}$, as in Fig. 3 (i). This interpretation is common, but it is not necessarily justified. Alternatively, it is possible that syntactic and/or semantic information directly modulates the values of the phonetic parameter in a single, non-hierarchical prosodic organization, as in Fig. 3 (ii).

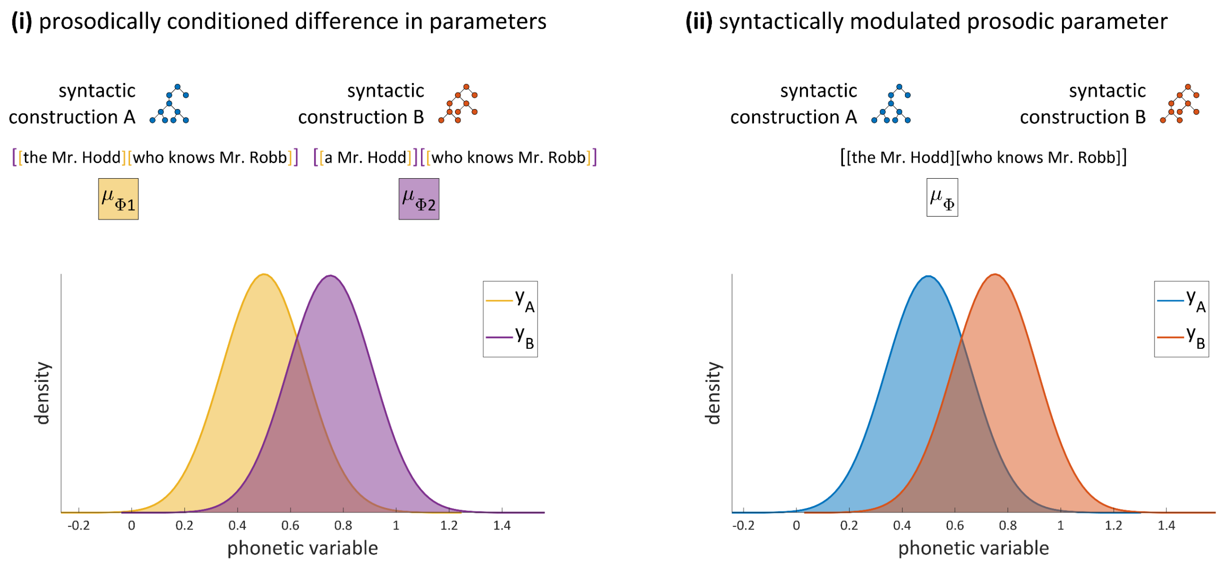

Fig. 3. A difference in distributions of a phonetic variable can be interpreted as a consequence of different prosodic phrase structures (i), or as a syntactically conditioned difference in a parameter associated with a single prosodic organization (ii).

Although many studies use paradigmatic comparisons to infer differences in hierarchical prosodic organization, their conclusions always hinge on the *a priori* assumption that such hierarchical organization exists, as in Fig. 3 (i). The alternative in Fig. 3 (ii) generally does not seem to be considered: syntactic/semantic information may directly modulate the value of a phonetic parameter associated with a non-hierarchical prosodic organization which is shared by both contexts. Either of these alternatives can generate the same pattern of phonetic variable distributions. Furthermore, this ambiguity holds whether the syntax-prosody mapping is deterministic or probabilistic: under reasonable assumptions, a probabilistic mapping would merely make the distributions associated with syntactic contexts A and B more similar. Note that a deterministic mapping can be viewed as a limiting case of a probabilistic mapping where the probabilities of prosodic structures conditioned on syntactic contexts are either 1 or 0.

Thus, analyses which rely on paradigmatic comparisons—i.e. phonetic variable differences associated with different syntactic/semantic contexts—do not provide strong evidence for hierarchical prosodic organization. In order to find more compelling evidence, it is useful to analyze the effects of speech rate on phonetic variables. There are some important consequences of how speech rate is measured and transformed, which we discuss in Section 2.3. Before discussing these, we examine how a generic "speech rate measure" can be useful for detecting categorical differences in prosodic phrase types.

Consider that in the paradigmatic comparison above, the only independent variable is the categorical variable of syntactic context. Presumably, phonetic measurements are collected either with spontaneous or experimentally conditioned variation in speech rate, and it is reasonable to suppose that speech rate may influence those measures. In particular, it may be that categorically different prosodic organizations interact with speech rate in different ways. As shown in Fig. 4, it could be the case that there are significantly different relations between speech rate and the phonetic variable in syntactic contexts A and B, i.e. a rate × context interaction effect. The figure shows a case in which prosodic organizations $\Phi 1$ and $\Phi 2$ are associated with different linear rate-effect slopes, $b_{\Phi 1}$ and $b_{\Phi 2}$, which result in different relations between phonetic variable y and speech rate. We refer to this as a *paradigmatic-interactional* comparison, because it results from comparing interaction effects associated with different syntactic contexts.

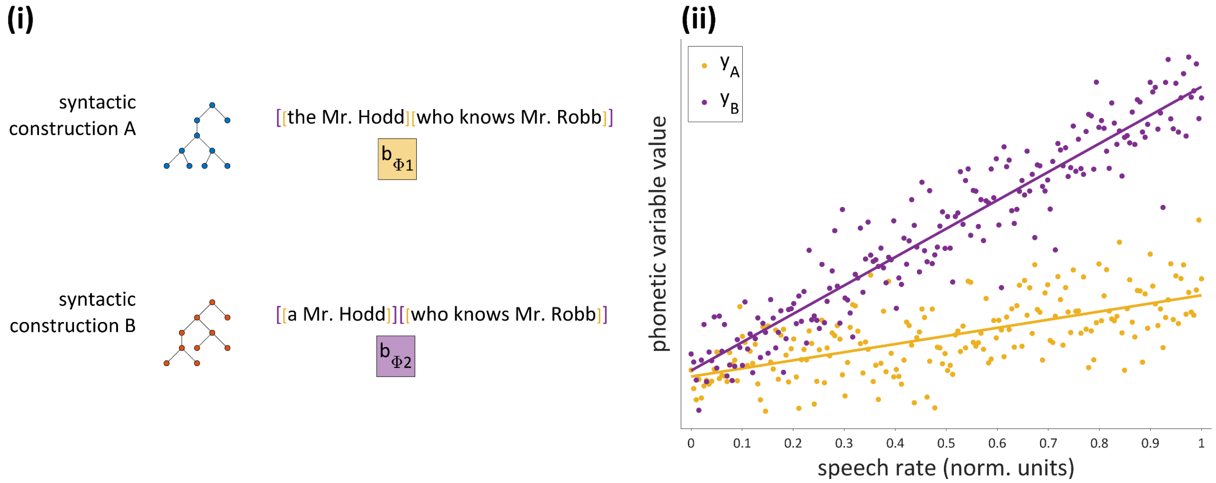

Fig. 4. An example of different speech rate effects in different syntactic contexts. (i) Syntactic contexts A and B map to prosodic structures $\Phi1$ and $\Phi2$, which are associated with different slopes for a speech rate effect, $\mathrm{b}_{\Phi1}$and $\mathrm{b}_{\Phi2}$. (ii) The effects of speech rate on a variable y in contexts A and B differ.

A paradigmatic-interactional effect is stronger evidence for a hierarchy of prosodic phrase categories than a paradigmatic effect. To infer that categorically different prosodic organizations are responsible for a paradigmatic-interactional effect such as the one in Fig. 4 (ii), an assumption is required. Specifically, it must be assumed that the influence of speech rate on phonetic variables is mediated by prosodic organization. Under that assumption, the simplest way to explain different effects of speech rate in different contexts is to posit that there are different prosodic structures/categories present in the contexts. However, it still remains possible that different effects of rate in contexts A and B are due directly to syntactic/semantic contextual differences. In other words, the same interpretative ambiguity that applies to the paradigmatic comparison applies here as well. We nonetheless believe that the inference that there are distinct prosodic phrases is stronger from a paradigmatic-interactional effect, because it is more reasonable to think of speech rate as an independent control parameter than as a syntactically conditioned parameter: it is uncontroversial that the same syntactic structure can be produced at a wide range of rates, and participants clearly have the ability to consciously control speech rate without special training.

The third type of evidence we consider is *within-context rate-conditioned mixture effects* (henceforth *within-context effects*), and we argue that such effects provide stronger evidence for categorical differences in prosodic organization than paradigmatic or paradigmatic-interactional comparisons. For within-context analyses, we test the predictions of the hypothesis that there is a speech rate-dependent mapping of a given syntactic structure to different prosodic structures. The intuition behind this analysis is that variation in speech rate may be associated with different phrasal organizations: a phrase boundary which in fast speech might be associated with a lower-level prosodic category (e.g. an ip) could be associated with a higher-level category (e.g. an IP) in slower speech. This predicts that dependent variables could exhibit distributions which show evidence of a mixture of categories, and that the probabilities of datapoints being assigned to each category should vary according to speech rate. The within-context analysis can provide stronger evidence than paradigmatic or paradigmatic-interactional analyses because it is conducted on each syntactic context separately, and thus does not suffer from the interpretative confounds discussed above—namely, ambiguity in whether effects are attributable to different prosodic structures or simply conditioned on syntactic contextual/informational differences.

Below we explain how within-context effects can be captured through regression mixture models. Consider that the within-context analysis hypothesizes a probabilistic relation between speech rate and prosodic phrasing: the likelihood that speakers may adopt some categorical prosodic organization depends on speech rate. From the perspective of statistical analyses, this entails that each observation includes not only a phonetic variable that is correlated with speech rate, but also a variable that indicates unobserved category membership. Regression mixture models (also known as mixtures of experts, or latent class variable models, and belonging to a class of models called finite mixture models; see Grün & Leisch, 2007; Jacobs et al., 1991; Kim et al., 2016; McLachlan & Peel, 2000) are well-suited for this circumstance. In the current context, we assume that there is an unobserved categorical variable that represents membership of an observation in one of two categories (labeled A or B below), and that the probability of category membership depends on speech rate, x. The logit function (log odds) is used to model these probabilities, as in (1) below. The interpretation of the parameters is more straightforward when viewing the inverse logit function, i.e. the logistic function in (2).

$$\log\left(\frac{\mathrm{p_A(x)}}{1-\mathrm{p_A(x)}}\right) = \alpha_0 + \alpha_1 \mathrm{x} \tag{1}$$

$$p_A(x) = \frac{1}{1+e^{-\alpha_1(x-\tau)}}, \quad \tau = \frac{-\alpha_0}{\alpha_1} \tag{2}$$

Some hypothetical examples of category probability functions are shown in Fig. 5A. The transition point in the speech rate dimension where category B becomes more likely than category A corresponds to $\tau = \frac{-\alpha_0}{\alpha_1}$, and the steepness of the transition corresponds to α_1. When α_1 is relatively large (Fig. 5A, green and blue lines), the categories are well separated by speech rate. However, when α_1 is relatively small (Fig. 5A, red line), the categories are not well separated by speech rate.

The regression mixture models that we consider allow for different regression coefficients for predicting the value of dependent variable $\mathrm{y_i}$ of observation i, according to the predicted category membership k, as in (3). k represents category, thus $k \in A, B$, and ϵ is normally distributed error with zero mean and category-specific standard deviation σ_k.

$$y_{i|k} = \beta_{0k} + \beta_{1k}x_i + \sigma_k\epsilon_i \tag{3}$$

Simulated data under two different conditions are shown in Fig. 5B and Fig. 5C. The simulations here were conducted with arbitrary, Gaussian-distributed dependent and independent variables. In Fig. 5B, the simulated data were drawn from two categories, according to the probability function shown in the bottom panel of Fig. 5B. In this simulation, category B had an intercept that was 2.0 units greater than category A, i.e. $\beta_{0A} = 0.0$, $\beta_{0B} = 2.0$. The effect of the speech rate measure (the independent variable) was the same for both categories, i.e. $\beta_{1A} = \beta_{1B} = 1.0$, and the standard deviations of the error terms were $\sigma_A = \sigma_B = 1.0$. Notice how in the vicinity of the transition point at $\tau = 0$ the simulation assigns some datapoints to category A and others to category B. The lines in Fig. 5B represent the category-specific simulation parameters and the filled interval is ± 1.0 σ.

A regression mixture model of the data in Fig. 5B is shown in Fig. 5B′. We used the R package *flexmix* (Grün & Leisch, 2007, 2008; Leisch, 2004) to fit the data (see Section 3.3 for more detail). The regression mixture model estimates not only the regression parameters of (3) but also the parameters of the logit function in (1). The model can be used to assign datapoints to one of the two categories (which is a form of clustering), based upon the estimated posterior probabilities of category membership for each data point. In the case of the model in Fig. 5B′, the reader can see that its parameter estimates provide a relatively close match to those that were used to generate the data. Moreover, the Akaike information criterion (AIC) of the mixture model is substantially less than the AIC of a simple linear regression (ΔAIC =

-14.6), and thus there is strong support for the mixture model over the linear model (see Section 3.3 for further detail on model comparisons).

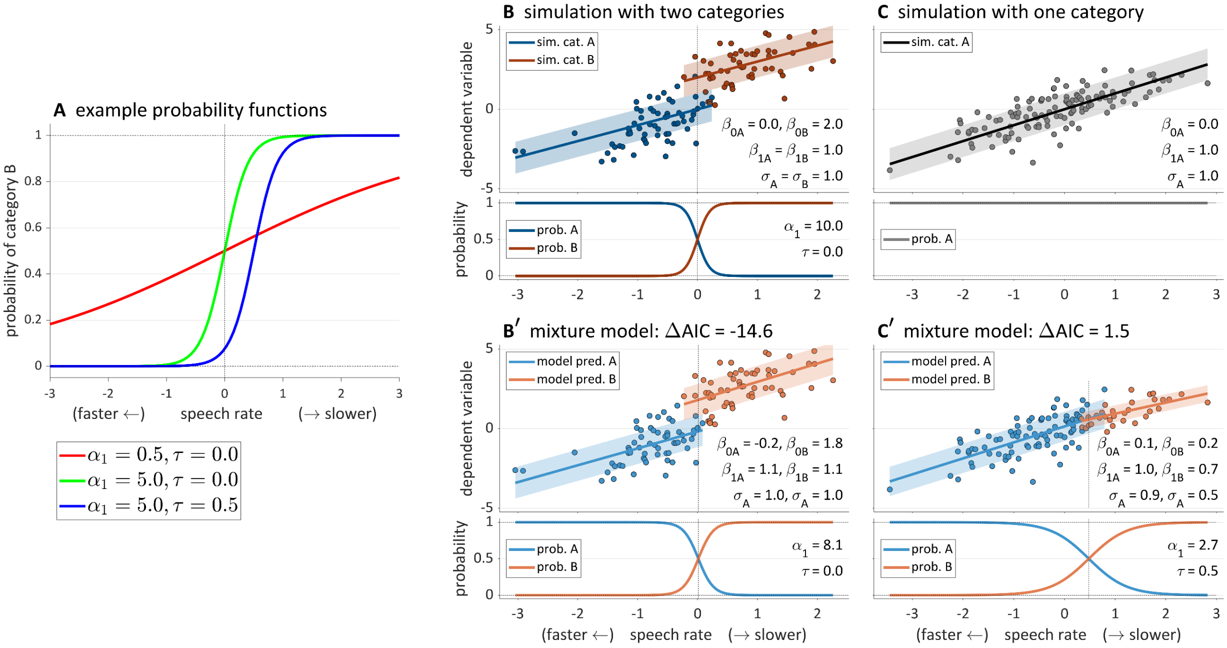

Fig. 5. Illustration of category probability functions and mixture models for simulated data. (A) Examples of probability functions as defined in (2) for several different combinations of parameters. (B) and (C): upper panels show simulated datapoints and parameters; lines are based on category-specific intercepts and slopes, shaded areas are ±1.0 standard deviation. Lower panels show simulation probability functions and parameters. (B′) and (C′): panels show estimated regression mixtures and probability functions, along with parameter values. Horizontal axes in all cases are a generic speech rate measure.

In Fig. 5C, the simulated data were drawn from a single category. The mixture model in Fig. 5C′ shows parameter estimates for two categories, but in this case, the AIC of the mixture model is not substantially less than the AIC of a simple linear regression model (in fact it is greater, i.e. ΔAIC = 1.5).

To verify whether the mixture models will consistently identify evidence for mixtures and consistently fail to identify evidence for mixtures when they are not present, we conducted the regression procedure 100 times for a variety of simulation parameters. Each simulation had 120 datapoints (the same number in our analyses below), and both independent and dependent variables were normally distributed. (See Appendix: Mixture model details for the sets of parameter values and the mixture analyses results.) We found that when the difference in category-specific intercepts was relatively large ($\beta_{0B} - \beta_{0A} = 2.0$, $\sigma_{A,B} = 1.0$) and the categories were sufficiently well separated ($\alpha_1 = 10.0$), the mixture was identified in 98.3% of cases (295/300). Conversely, when the simulation imposed no difference in category-specific parameters (which is the same as the case of a single category), evidence for the mixture was falsely detected in 2.9% of cases (26/900). We also found that there are a variety of circumstances in which the mixture model can fail to identify underlying categories. Specifically, this can occur when the difference in category intercepts is relatively small, when the categories are not sufficiently separated along the speech rate continuum, or when the category-specific slopes (i.e. speech rate effects) obscure the difference in category-specific intercepts —see Appendix for examples.

Note that in within-context analyses, we test only for mixtures of two categories, rather than three or more. This constraint was adopted in part to keep the analyses simpler, and also because more observations may be necessary to resolve more mixture components. Thus, the analyses do not provide evidence regarding whether speakers adopt three or more categorically different prosodic organizations within a given syntactic context. It is also important to point out that the analyses can also succeed in identifying two categories under the hypothesis that the rate-conditioned prosodic organization is deterministic. For example, in a deterministic rate-conditioning, the probability of category A would be 1 below some particular speech rate value, and above that value, the probability of A would be 0. In the mixture model, the deterministic mapping can be viewed as a limiting case where the growth rate parameter goes to infinity, which creates a step-like jump in the probability function.

To summarize the above discussion: we argue that there is a range of quality of evidence for the existence of categorical differences in prosodic organization. Paradigmatic comparisons provide relatively weak evidence, because it is ambiguous whether differences are conditioned by prosodic organization or by syntactic/semantic context. Paradigmatic-interactional comparisons provide somewhat stronger evidence, because speech rate effects are more likely to be mediated by prosodic structure; however, the same ambiguity that exists for paradigmatic comparisons also applies to paradigmatic-interactional comparisons. Within-context mixture models provide the strongest evidence. Overall, the incorporation of speech rate as an independent variable helps to draw inferences about prosodic organization; it also eliminates contextual factors which make interpretations of effects ambiguous. Nonetheless, it is not necessarily possible to detect any arbitrary mixture of categories, and for reasons we discuss in Section 5, it may not be possible to differentiate within-context mixture of categories from scale attenuation effects. Thus, to our knowledge, there exists no unassailable form of phonetic evidence for the existence of distinct prosodic categories.

2.3. Elicitation and measurement of variation in speech rate

Speech rate plays an important role in detecting evidence for differences in prosodic organization. However, there is no unique definition of speech rate; instead, there are many different ways that rate can be measured and expressed quantitatively. To avoid confusion, we use the label *rate measure* generically to refer to any quantity which may correspond to our subjective impression of how quickly a participant is speaking, regardless of how that quantity is expressed.

Rate measures are typically derived from counting a number of similar events that occur and dividing that count by the period of time over which the count is taken (i.e. the measurement period). For example, we might express rate as a count of how many syllables, words, or sentences, etc. occur per second. In our experiment, the numbers of syllables (14), words (10), and sentences (1) per trial are always the same, and only the measurement period (i.e. response duration) varies. The corresponding rate measures (i.e. syllables/s, words/s, sentences/s) are simply multiplicatively scaled versions of each other, and thus the comparisons we make between regression models do not depend on which event is chosen as the basis of the rate measure.

Yet, there is a common transformation of rate measures that may have important consequences for analyses of categorically distinct prosodic organization. Specifically, any rate measure expressed in units of events/s can be transformed via the multiplicative inverse function (i.e. reciprocal) to an average period of time per event, i.e. s/event. For example, we might express the rate for a given trial as the average syllable duration, the average word duration, or even the duration of the sentence itself. Henceforth we will refer to such measures as durations or *inverse rates*, in contrast with measures based on counts per second, which we refer to as *proper rates*, or frequencies.

The above distinction is important because proper and inverse rates are nonlinearly related, and hence when used as independent variables they may lead to different results in regression models. This point is illustrated in Fig. 6 with simulated data drawn from two categories where category membership depended on inverse rate (left column) or proper rate (right column). In both cases, the relation observed between the independent variable of speech rate and a dependent variable differs depending on whether the rate measure is expressed as a proper rate (bottom row) or inverse rate (top row); in other words, the two figures within each column have different functional relations according to how the rate measure is quantitatively expressed. These differences will have consequences for linear regression models and mixtures of regressions, and hence it is important to bear in mind that interpreting the results may depend on how the rate variable is expressed. To reinforce understanding of the relation between inverse and proper rate measures, two datapoints in each simulated dataset (i.e. for each column of the figure) are shown with colored circles.

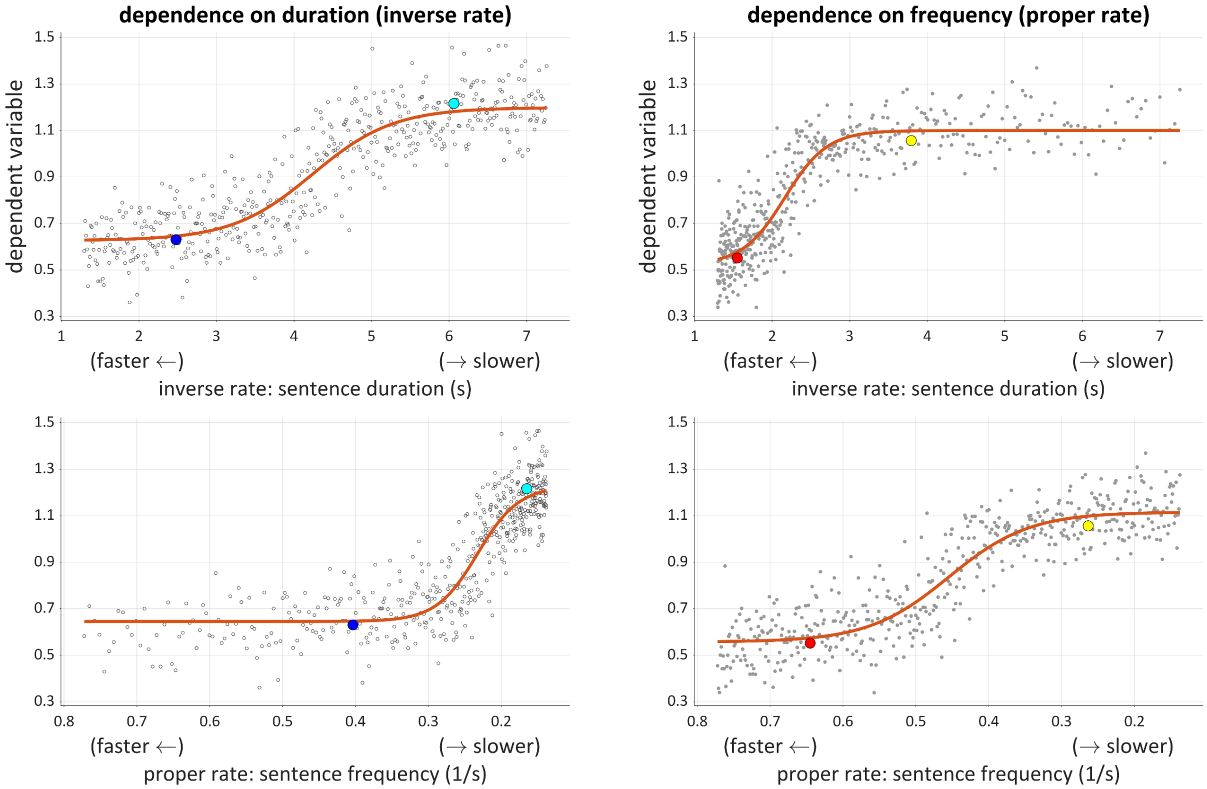

Fig. 6. Illustration of effects of how rate measures are expressed. Left panels: data were generated with dependence on inverse rate; right panels: data were generated with dependence on proper rate. Top panels: rate measure is expressed as a duration; bottom panels: rate measure is expressed as a frequency. In both left and right panels, the data show different patterns according to the choice of rate measure. The orientation of proper rate coordinates is reversed so that slower rates are to the right. Smoothing spline fits of data are provided to illustrate differences in functional relationships. The colored circles mark the same trials (identical dependent values) expressed in different rate measures.

To our knowledge, there is no established theoretical basis for choosing inverse or proper rates as independent variables. A reasonable theoretical motivation might be available if we better understood how participants control their rate of speech. In the absence of this understanding, we opted to conduct analyses with both inverse and proper rates, and we assess the consequences of the choice of rate measures. However, there does appear to be an empirical motivation for preferring inverse rate over proper rate in our experiment: an analysis of rate measures by participant (see Section 3.3) shows that distributions of inverse rates (specifically sentence durations) are less skewed and more uniform (have lower kurtosis) than distributions of proper rates.

There are two additional points to make regarding rate measures here. First, the rate measures we employ as predictors can be described as *effective* or *empirical* measures, in the sense that they are derived from the productions of participants; these contrast with *target rates*, which are rate values along a continuum that we employed to elicit variation in rate (see below and Section 3.1 for more detail). Second, the rate measures we use are relatively global, in the sense that they represent the maximal period of time we have available for measurement—the entire response. It is important to recognize that speech rate may also be modulated locally, in particular at prosodic boundaries, and the local rate modulations may affect dependent variables in a different way from global modulations. Indeed, some of the dependent variables we examine can be interpreted as reflecting local rate modulations. Unless otherwise noted, references to speech rate in this manuscript connote a global speech rate.

The method we employed for eliciting variation in speech rate is also important. In order for variation in speech rate to be maximally useful for regression analyses, the variation should be sufficiently large and should be relatively uniformly distributed. To that end, typical approaches to elicit variation in rate may not be adequate. Most studies which explicitly elicit rate variation do so by instructing participants to speak at qualitatively different rates often described as “slow” vs. “medium” vs. “fast”, or perhaps by instructing them to speak “carefully” vs. “casually”. While there are several variants of the qualitative descriptors that are used in instructions, what such approaches have in common is that they impose a handful of *ad hoc* categories (typically two or three) on what is more fundamentally a continuum of rates—we thus refer to these approaches as a *categorical rate instructions*.

There are several problems with categorical rate instructions. One is that different participants may adopt drastically different interpretations of the categories or may adopt very different mappings of the rate categories to effective rates. Another is that participants may produce multimodal distributions of rates, as opposed to a more uniform distribution. These are problems because they make the estimation of continuous rate effects less robust.

To avoid the above issues with categorical rate instructions, the experimental design elicited non-categorical variation in rate by using a moving visual analog cue for rate. Participants saw the cue before each response and were instructed to base the speed of their response upon the speed of the cue. (Note that participants did not produce the utterance during the cue; instead, the cue iconically represented the target rate of the subsequent response). Ten different cue rates were used, and the cue rates were varied randomly from trial to trial; this prevented participants from identifying distinct rate categories.

2.4. Hypotheses and predictions

In order to assess evidence for a hierarchical organization of prosodic phrase categories, we recorded acoustic and articulatory data from productions of RRCs and NRRCs, with variation in speech rate elicited as described above. For paradigmatic and paradigmatic-interactional comparisons, we conducted linear regressions with syntactic context and speech rate as predictors as well as their interaction, within each variable at each boundary. To conduct the within-context mixture model analyses, a regression mixture model was fit to each variable at each boundary in each syntactic context; the Akaike Information Criterion (AIC) was used to assess whether the model provides a substantially better fit than a simple linear model, and several additional criteria were imposed to identify mixtures which are consistent with the hypothesis that there are two distinct phrasal categories (see Section 3.3 for more detail). All analyses in this paper were conducted within participant because we suspect that participants do not necessarily employ the same syntax-prosody mapping, and we do not aim to draw inferences regarding population-level parameter estimates. The following hypotheses are tested:

Hyp. 1. *Syntactically-conditioned variation in prosodic organization*: differences in hierarchical prosodic phrase organization are associated with different syntactic structures. This hypothesis predicts that in linear regressions, models with relative clause type as a factor will provide significant improvement over models without relative clause type (a paradigmatic effect); or models with a clause type-speech rate interaction will provide significant improvement over models lacking this term (a paradigmatic-interactional effect).

Hyp. 2. *Rate-conditioned variation in prosodic organization*: hierarchical prosodic organization changes with speech rate, within a given syntactic context. This hypothesis predicts that two-category mixture models of the relation between speech rate and dependent variables will offer substantial improvement over linear models and that the categories identified in mixture models should be well-separated along the speech rate continuum.

Due to the exploratory nature of the analysis, we do not make specific predictions regarding which variables are more or less likely to exhibit syntactically-conditioned and/or rate-conditioned effects. In principle, all of the phonetic correlates of prosodic organization discussed in Section 2.1 might show the effects predicted by the hypotheses. Yet there are many factors beyond syntactic organization that might influence the measurements, so it is not anticipated that all variables or even a majority will show the predicted effects.

Furthermore, as explained above, the analyses used for Hyp. 1 (i.e. paradigmatic and paradigmatic-interactional comparisons) provide somewhat weaker evidence than the within-context analyses of Hyp. 2, because the former presuppose that syntactic/semantic differences cannot be directly responsible for prosodic variation. This raises the question: why conduct the paradigmatic and paradigmatic-interactional analyses at all? We believe that conducting these analyses has some value precisely because those results can be subjected to criticism; moreover, since such analyses are more standard, they may serve as a useful basis of comparison. Although Hyp. 2 provides stronger evidence for categorical differences in prosodic organization, the within-context analyses of this hypothesis still require a number of assumptions about the relations between rate, prosodic organization, and measurements, as discussed in Sections 2.2 and 2.3; namely, it is assumed that there are potentially two different phrasal organizations (as opposed to three or more), and that these organizations are at least moderately conditioned on speech rate. It is unlikely that these assumptions hold in all cases, and this must be taken into account in interpreting the results.

In addition to testing the above hypotheses, we conduct an analysis of the consequences of expressing the rate measure as a duration (inverse rate) or as a frequency (proper rate). We do not have specific predictions regarding how this choice of independent variable will affect the analyses for Hyp. 1 and 2. This analysis is nonetheless important because it may inform our understanding of whether inverse or proper rates are more appropriate independent variables, and it may reveal differences in the interaction between speech rate and dependent variables which otherwise would not be apparent.

3. Methods

3.1. Participants and task

Twelve native speakers of English (6M, 6F) with no speech or hearing disorders participated in the experiment. For six of these (3M, 3F), only acoustic data were collected (acoustic-only sessions); for the other six, articulatory data were collected in addition to acoustic data (articulatory sessions). The experimental task was identical in both cases. In articulatory sessions, participants were seated about 1.5 m from a computer monitor in a quiet room with a shotgun microphone positioned 1.5 m from their mouth; in acoustic sessions, participants were seated in front of a computer monitor in a sound-attenuated booth and wore a condenser microphone (AKG C520 headset). There were six blocks of 40 trials in each experimental session. One type of RC was produced throughout a block, and the blocks alternated between the two types of RCs.

Participants followed the procedure in Fig. 7. In each trial, participants first saw two sentences in sequence on the screen. The first sentence (Fig. 7 (1)) provided context to the second sentence (Fig. 7 (2)), which was the target sentence in the experiment. (cf. An example sentence pair for the RRC context is introduced in Table 1.) The purpose of the context was to draw attention to the semantic/pragmatic differences between the two types of RCs, which otherwise would have been distinguished just with the presence of commas in the target sentence. For example, the context “There are two Mr. Hodds. Only one knows Mr. Robb” picks out one Mr. Hodd in particular, favoring the RRC interpretation of the target sentence. Conversely, the context “There is one Mr. Hodd. He knows Mr. Robb” facilitates the NRRC interpretation of the target as the second sentence (He knows Mr. Robb) simply gives extra information about the referent (Mr. Hodd). In addition to providing the context sentence, determiners in the beginning of the target sentence (*A, The*) served the same purpose of cueing participants on the difference between NRRC and RRC. The determiner “the” favors the RRC interpretation because it presupposes the existence of more than one person named “Mr. Hodd”. No participants reported any difficulty in interpreting the meanings of the two types of sentences.

Participants were instructed to read both the context and target sentences silently when they first appeared (Fig. 7 (1) and (2)). After 1.5 seconds, the visual rate cue appeared, which is the red box in Fig. 7 (3). This cue moved from left to right across the screen. There were 10 different speeds of the cue used throughout the experiment. These are *target rates*, which were defined according to the period of time it takes for the cue to move across the screen. The periods were equally spaced from 0.8 s to 4.1 s. The endpoints of this range were based on observed durations of the target sentence in pilot data, which were conducted on two native speakers of English. Note that this cue can be considered an “analog” stimulus because participants were not able to identify the 10 unique rates; rather their subjective impression was that the cue rate varied on a continuous scale. Participants were instructed that when the cue disappears and the gray box changes to green, they should produce the target sentence in a way that reflects variation in the speed of the visual cue. Crucially, participants were told to wait until the box becomes green before initiating their response. Participants therefore did not match their production to the period of time in which the cue was visible, but instead, they varied the speed of their production based upon their impression of how fast or slow the cued moved across the screen. Thus, the rate cue was used to indirectly elicit a wide variation in rate rather than imposing a specific timeframe in production.

The speed of the rate cue was varied randomly from trial to trial by selecting one from the set of 10 target rates. Each rate appeared in the same number of times in the experimental session. Note that the steps of the rate continuum, namely the target rates, were not used as predictors in analyses. Instead, we used measures of *effective* speech rate, which are based on the productions of the participants. The reason for this is that measures of effective speech rate better reflect the continuous nature of speech rate variation that was elicited from participants and better account for between-participant variation in the ranges and scales of speech rate.

To check that participants did vary their rates, we examined the range of inverse rates (durations of the produced target sentences) within each participant. The minimum range for any participant was 1.67 s, the maximum range was 5.36 s, and the average range across participants was 2.41 s. This shows that the rate cueing method was successful in eliciting a wide variation of speech rate without using categorical rate instructions such as *speak fast or slow*.

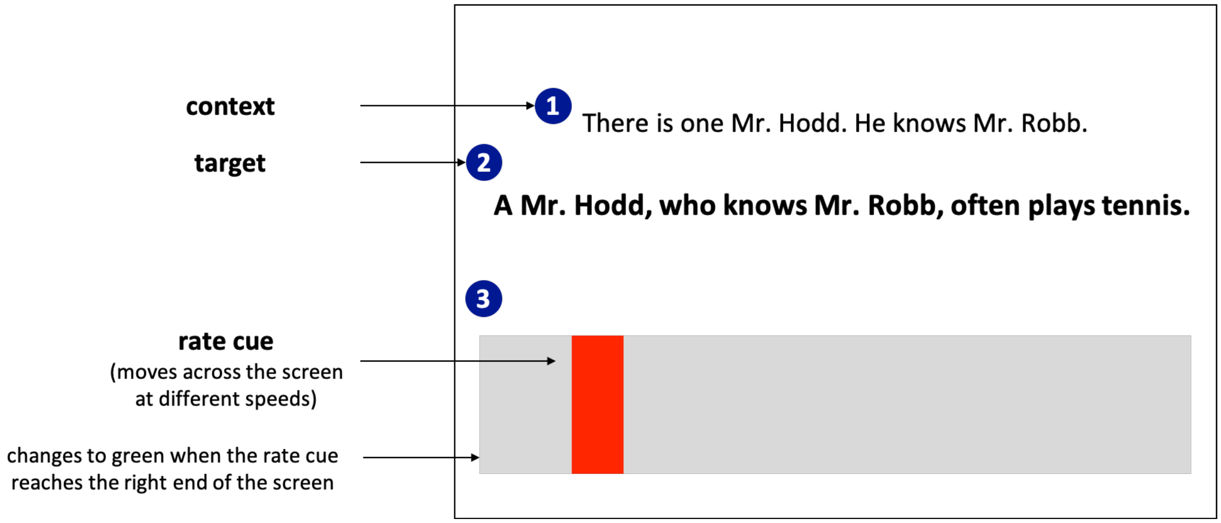
context
target
rate cue
(moves across the screen
at different speeds)
changes to green when the rate cue
reaches the right end of the screen
1
There is one Mr. Hodd. He knows Mr. Robb.
2
A Mr. Hodd, who knows Mr. Robb, often plays tennis.
3

Fig. 7. Presentation of a single trial. Participants were instructed to read the context and target sentences silently when they appeared in sequence (1 and 2). The rate cue (red box) then showed up and moved across the screen at different speeds (3). When the cue reached at the end of the screen and the gray box changed to green (not shown in the figure), participants read the target sentence (bold) in a way that reflected variation in the speed of the rate cue.

The target words in the experiment were the names that follow the honorific “Mr.” All the names were monosyllabic with a CVC form. The onset consonants of these targets were /h/, /r/, or /l/, and the codas were /b/ or /d/. We restricted the set of coda consonants to these two segments in order to facilitate articulographic analysis of boundary effects. Note that each participant produced 120 /b/-final and 120 /d/-final forms at each boundary. The vowel /a/ was used in all cases. In order to prevent participants from putting emphatic focus on the names, they were explicitly instructed not to emphasize them. In addition, participants performed sixteen practice trials under experimenter supervision before beginning the experiment. If they put focus on the names, they were corrected by the experimenter.

3.2. Data collection and processing

In both articulatory and acoustic-only sessions, acoustic data were collected at 22050 Hz. Acoustic segmentations were conducted using Kaldi (Povey et al., 2011). For each participant, six trials were manually labeled and used to train monophone HMMs. Then, a forced alignment was conducted for all remaining trials. The alignments of all trials were manually inspected and corrected. Acoustic durations of the coda, vowel, and rime portions of the target words were obtained from the segmented data.

F0 data were extracted using Praat. We first identified a participant-specific F0 range with the first 20 trials of each participant. Specifically, we collected all F0 values of the first 20 trials, removed outliers, and calculated the range of the remaining F0 values, which were also sanity checked considering the gender of the participant. With this participant-specific F0 range, F0 timeseries were extracted in Praat with a timestep of 5 ms using auto-correlation method. The F0 timeseries were used to obtain the mean and maximum F0 of the vowels before and after the pre-/post-relative clause boundaries (henceforth, we refer to these boundaries as B1 and B2). Specifically, the F0 measurements were conducted on /a/ in the first target word (pre B1), /ʊ/ in *who* (post B1), /a/ in the second target word (pre B2), and /a/ in *often* (post B2). Note that we only used the middle third of the F0 frames of each vowel in measuring these variables in order to avoid microprosodic effects on F0 at vowel margins. In addition to these F0 measures, we also calculated F0 differences between the vowels immediately preceding and following prosodic boundaries (post B1 – pre B1, post B2 – pre B2), since pitch reset is likely to occur at the boundaries.

Articulatory data were collected with an NDI Wave Electromagnetic Articulograph (EMA) with a sampling rate of 400 Hz. Articulator sensors were located mid-sagitally on the upper lip (UL), lower lip (LL), gum below the lower incisors (JAW), and tongue tip (TT, about 1 cm from tongue apex) and body (TB, 4-5 cm posterior from TT). Reference sensors for head movement correction were located on the nasion and left and right mastoid processes.

Articulatory data were processed as follows. First, reference and articulator sensors were filtered at 5 and 10 Hz respectively, using 3rd order lowpass Butterworth filters. Then, head movement was corrected by transforming each frame of data such that the reference sensors were located at a fixed position. The horizontal and vertical coordinates of the articulator sensors were then resampled to 1000 Hz to allow for more precise identification of gestural landmarks. A lip aperture signal (LA) was defined by calculating the Euclidean distance between the UL and LL sensors. Fig. 8 shows the kinematic landmarks along with the dependent measures for the articulatory analyses. Based on the acoustic segmentation, relevant velocity extrema were detected for consonantal closure and release gestures of the coda of the target words at each boundary (green dots in Fig. 8). Gestural onsets and targets were then identified in relation to velocity extrema. Specifically, the gestural onsets were the time points where the velocity signal first rises above 20% of the peak velocity, and the targets were the time points where the signal first falls below 20% of the maximum velocity, which are shown as red and blue asterisks in Fig. 8. The target words in the experiment ended either in /b/ or /d/; thus, for the words that ended with /b/, the LA signal was analyzed, while for the words that ended with /d/, the vertical position of the TT sensor was analyzed. For articulatory analyses, we measured onset-to-release intervals (closure + plateau durations) and movement durations at the release phase of the consonant. Furthermore, amplitudes of closure and release movements and peak velocities were measured at each boundary.

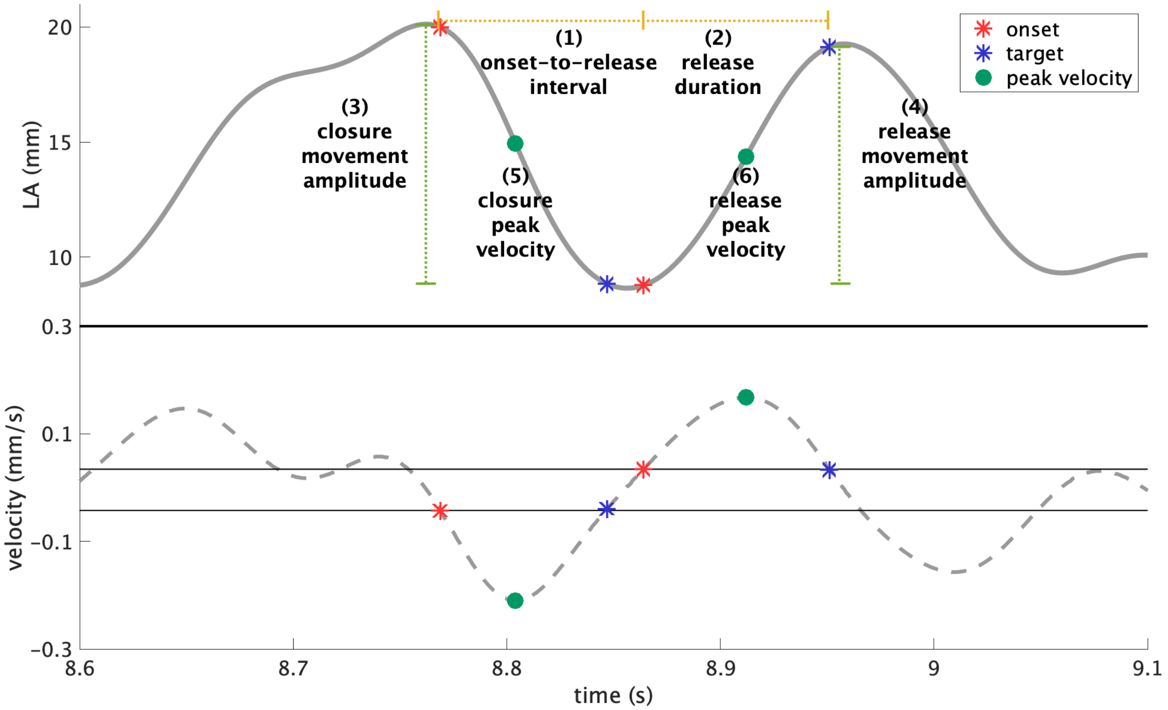

Fig. 8. An example showing an LA trajectory (top panel) and the accompanying velocity trajectory (bottom panel) with markings of kinematic landmarks and dependent articulatory variables. For each boundary, the onset-to-release intervals (closure + plateau durations), movement durations at the release phase of the coda, and the amplitudes and peak velocities of closure and release movements were measured.

To characterize local speech rate slowing or pausing at prosodic boundaries, we used a hybrid articulatory/acoustic measure instead of pause durations. Pause durations are somewhat problematic because it can be hard to distinguish pauses from periods of low acoustic intensity associated with the often devoiced codas of the target words; moreover, pause durations are not defined on trials in which no silent interval is detected, which makes them less suitable for use in regression analyses. To circumvent these issues, we defined a hybrid articulatory/acoustic measure, the trans-boundary interval (TBI), which is well-defined on all trials. Specifically, the TBI is the period of time from the target achievement of the pre-boundary coda closure (measured articulatorily) to the acoustic onset of the post-boundary vowel – this vowel onset closely corresponds to the onset of voicing. For example, consider the utterance “A Mr. Hodd, who knows Mr. Robb, often plays tennis”. Here the TBI of B1 is the time from the target achievement of the alveolar closure in *Hodd* to the start of the vowel [ʊ] in *who*; likewise, the TBI of B2 is the time from the target achievement of the bilabial closure in *Robb* to the start of the vowel [a] in *often*. The TBI can thus be readily interpreted as a measure of local speech rate slowing/pausing associated with a phrase boundary.

A total of 240 trials were collected for each of the 12 participants. For articulatory sessions (1440 trials), a total of 183 trials (12.7%) were discarded: 127 trials due to speech errors, disfluencies, or problems in data collection, and 56 trials due to problems in detecting landmarks. For acoustic-only sessions (1440 trials), 50 trials (3.5%) were discarded for speech errors, disfluencies, and data collection problems. This left 2647 trials out of 2880 trials in total (91.9%). For all measurements, a mixed effects linear regression with response sentence duration (inverse rate) as a fixed effect and participant as a random intercept was conducted at each boundary in order to identify outliers. Datapoints whose standardized absolute residuals were larger than 2.326 (99%) were excluded from subsequent analyses.

3.3. *Data analysis*

As explained in Section 2.3, the rate measures are independent variables in our analyses. Rate measures are quantitative indices of how quickly speakers are speaking. We use effective rate measures (measures based on participant responses) rather than target rates (visual analog cue rates), because the effective measures more precisely characterize the global speech rates that participants employed. Furthermore, we used two types of rate measures as predictors – inverse rate (duration) and proper rate (frequency). Inverse rate was defined as the duration of the produced sentence and has units of seconds; proper rate was defined as the reciprocal of inverse rate and has units of sentences/s.

For each of the three analyses we conducted (paradigmatic, paradigmatic-interactional, and within-context), the analysis was conducted both with inverse and proper rates. Fig. 9 shows distributions of inverse rate and proper rate within each participant as well as skewness and kurtosis measures of each distribution. The analyses show that effective inverse rate tends to be more uniformly distributed (lower kurtosis) and also more symmetric (less skewed) than the proper rate. Note that kurtosis values are expressed relative to the kurtosis of a normal distribution in Fig. 9. Since inverse rates have more desirable distributional properties, Section 4 presents the results of these analyses first, and then compares them to the results of the analyses conducted with proper rate as an independent variable.

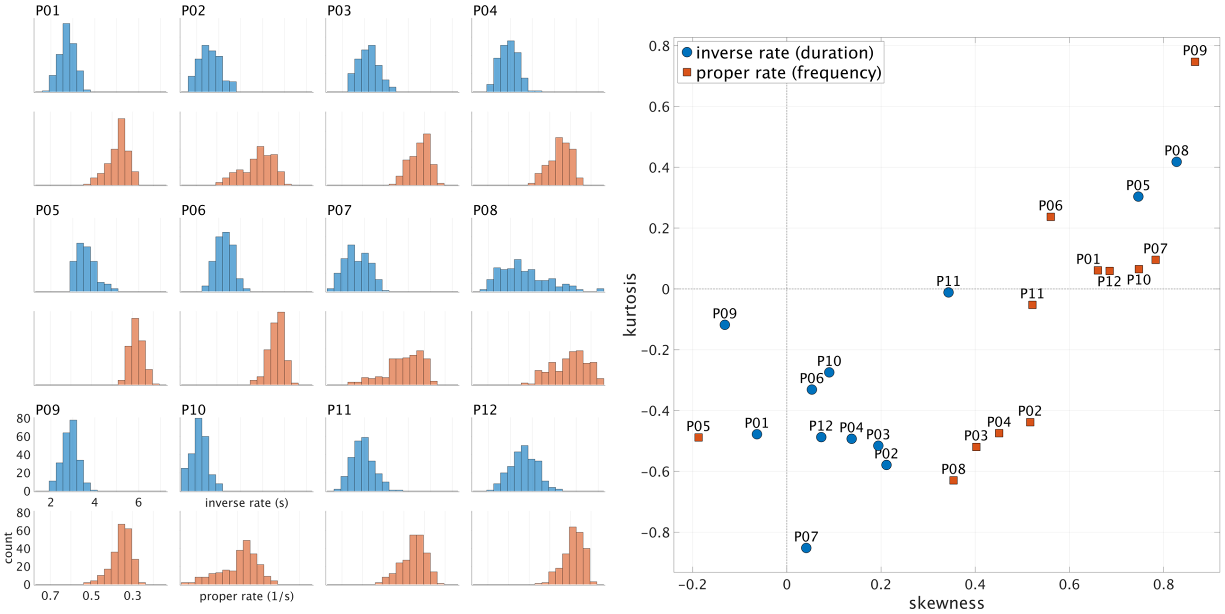

Fig. 9. Distributions of inverse rates and proper rates within each participant. The left panel shows the histograms of inverse rates (blue) and proper rates (orange) for all participants, and the right panel plots kurtosis vs. skewness for each participant/rate measure (blue circles: inverse rates; orange squares: proper rates). Kurtosis values are expressed relative to the kurtosis of a normal distribution. The orientation of proper rate coordinates is reversed in the left panel.

All analyses were conducted within participant. We do not conduct across-participant analyses for several reasons. First, we do not expect that all participants will necessarily employ the same boundary strengths or prosodic phrase organization in the same syntactic contexts, nor do we expect that the effects of rate on dependent variables and categories will be similar across participants. Hence a mixed effects model in which there are random terms for participants would be inappropriate for analyses. Second, there are substantial differences in the distributions of speech rates between participants (see Fig. 9); it is thus unclear whether the rate measures are commensurate between participants, and this raises complications for any analysis which combines data from all participants. Third, the goal of our analysis is not to draw statistical inferences regarding population level fixed effects; in other words, we are not specifically concerned with the extent to which the RRC vs. NRRC contrast influences various measures or in identifying "an effect of speech rate" on those measures; instead, we are concerned primarily with the amount and quality of evidence for distinct prosodic categories, and how that evidence may depend on the expression of speech rate.

Paradigmatic and paradigmatic-interactional analyses were conducted with a linear regression model. For each measurement at each boundary within each participant, we conducted a step-wise regression procedure, in which we first tested for an interaction effect between speech rate and RC type (NRRC/RRC), and if no such interaction was found, we tested for a main effect of RC type. The full model included RC type, speech rate, RC type × speech rate interaction, and place of constriction. Place of constriction was only included as a covariate for articulatory variables and acoustic durations, but not for F0-related variables. We used log-likelihood ratio tests to compare the models. If the full model offered a significant improvement over the model without the RC type × speech rate interaction, we counted it as an instance of a paradigmatic-interactional effect. If the full model did not significantly improve the fit, we subsequently removed the interaction term and compared a model with three main effects (i.e. RC type, speech rate, place of constriction) to a model without the RC type effect. If the former showed a significant improvement over the latter, it was counted as an instance of a paradigmatic effect. Because paradigmatic comparisons were only assessed in the absence of paradigmatic-interactional effects, we also present the combined proportions of paradigmatic and paradigmatic-interactional effects.

To test for within-context rate-conditioned mixtures, we used the R package *flexmix* (Grün & Leisch, 2007, 2008; Leisch, 2004), which uses the Expectation-minimization (EM) algorithm to fit finite mixtures of regressions. The category-membership model we specified was a binomial logit model with speech rate as a predictor (see (1) above), and the regression had speech rate as a predictor with category-specific intercepts, slopes, and error variances (see (3) above). The initial category memberships supplied to the algorithm were distributed such that category A was associated with datapoints in the lower half of the range of speech rates, and conversely, category B was associated with datapoints in upper half of the range of rates; in the analyses, we refer to category A and B as category F (category associated with fast rate) and S (associated with slow rate) respectively.

In order for the estimated models to be counted as evidence for the presence of two categories, the EM algorithm had to converge over 10000 iterations and the AIC of the mixture model had to be 2 less than the AIC of the simple linear model (see Burnham & Anderson, 2004). In addition, we imposed several criteria for the mixture model to be considered as evidence for two categories: (i) at least a third of the datapoints had to belong to the category that had fewer members; (ii) the transition point in the estimated category membership function (i.e. the location of 0.5 probability) had to occur within the range of speech rates observed; and (iii) at both extremes of the empirically observed rate continuum, the difference in category probabilities had to be greater than 0.80, reflecting the condition that each of the two categories should predominate at opposite ends of the rate continuum; in other words, the datapoint at the slowest end should at least have a probability of 0.9 to be associated with the slow-rate category (category S), and likewise, the datapoint at the fastest end should at least have a probability of 0.9 to be associated with the fast-rate category (category F) (cf. the latter is equal to having a probability of less than 0.1 to be associated with the slow-rate category). These criteria implement adherence to the hypothesis that categorical differences in phrasal organization are conditioned on speech rate. In inverse rate analyses, although 315 mixture models were substantially better fits than the linear regression models, 194 of those (61.6%) did not pass the three criteria. In proper rate analyses, 298 mixture models were substantially better fits than the liner regression models, but 191 cases (64.1%) did not pass the additional criteria.

For the within-context mixture regressions analyses, the effects of coda place (i.e. LAB vs. COR) were partialled out of the articulatory and acoustic variables. To accomplish this, the place effect coefficient was estimated with a simple linear regression and then subtracted from the data before the mixture model parameters were estimated. Place effects were subtracted prior to the analysis rather than included in the model, in order to simplify the model and facilitate convergence. Place effects were not subtracted from F0-related variables.

The proportions of within-context effects were compared across boundaries, phonetic variables, and participants. We also examined mixture model fits of the cases that showed significantly better fit in the mixture model over the linear one. In analyzing the model fits, we compared the magnitudes of the slopes of category F and category S components and determined which category was more responsive to rate variation (i.e. compare $|\beta_{1F}|$ vs. $|\beta_{1S}|$). Furthermore, we examined whether the slope of the more responsive category is positive or negative (e.g. in fast-rate responsive pattern, $\beta_{1F} > 0$ vs. $\beta_{1F} < 0$). Note that in the proper rate analyses, we examined the negative value of the slope of each category; in the proper rate coordinates, the slower rates are on the left side of the x-axis and faster rates are on the right side (e.g. 0.3 sentences/s is slower than 0.6 sentences/s), which is the opposite of the inverse rate coordinates (e.g. 1.5 sec is faster than 3 sec). Thus, to facilitate comparison between the analyses of the two rate measures, $-\beta_{1F}$ and $-\beta_{1S}$ were examined in the proper rate analyses.

The within-context analyses were conducted separately for each syntactic context, measurement, boundary, and participant, which resulted in a total of 588 cases (i.e. 144 articulatory measures + 144 acoustic durations + 24 TBIs + 276 F0 measures). On the other hand, paradigmatic and paradigmatic-interactional comparisons were conducted in each measurement, boundary, and participant, and a total of 294 cases were examined. Note that we could only investigate TBIs in the data from six participants in the articulatory sessions as TBIs were defined with articulatory and acoustic measures. In addition, 12 F0 measures from one participant (P05) were excluded from the analyses because that participant produced a lot of creaks particularly at B2.

4. Results

Overall, we did not find extensive phonetic evidence for the existence of multiple levels of prosodic phrase categories. The strongest form of evidence, within-context mixture effects, was observed in 20.6% (121/588) and 18.2% of cases (107/588) in inverse rate and proper rate analyses respectively. In both analyses, effects were generally more prevalent at B2 compared to B1. In addition, we found relatively high proportion of within-context effects in variables that represent boundary-localized slowing and gestural overlap, such as the TBI variable, articulatory movement amplitude of the release gesture, and coda duration, especially at B2. Paradigmatic-interactional and paradigmatic effects were more widespread than within-context mixture effects, as roughly 40% of cases exhibited these effects. However, they do not provide compelling evidence for hierarchical phrasal organization because they can be interpreted as direct effects of syntactic/semantic differences. The comparison between the analyses conducted with inverse rate and proper rate found that the mixture model fits of acoustic durations were dependent on the choice of rate measure.

As the analyses with inverse rate and proper rate did not show a large difference and inverse rates showed more desirable distributional properties (see Section 3.3), we first provide results on the analyses with inverse rate as an independent variable. Section 4.1 presents the results of paradigmatic and paradigmatic-interactional comparisons of the inverse rate analyses, and Section 4.2 presents the results of within-context mixture analyses conducted with inverse rate. In Section 4.3, we examine how the choice of rate measures affects our analyses. We therefore compare the results in Sections 4.1 and 4.2 with the analyses results conducted with proper rate as an independent variable.

4.1. Paradigmatic and paradigmatic-interactional analyses

There were a fair amount of cases in which significant paradigmatic or paradigmatic-interactional differences were observed, which supports Hyp. 1; yet we argue that this finding does not provide evidence for a hierarchy of prosodic phrases due to the reasons we discussed earlier (see Section 2.2). A total of 294 cases were examined, and 125 of them (42.5%) showed either paradigmatic or paradigmatic-interactional differences: specifically, 78 paradigmatic and 47 paradigmatic-interactional differences were identified. Among different phonetic variables, the strongest effects were found in TBIs, which showed significant differences in more than 60% of the data (see Table 3).

Table 3. Number of cases that showed significant paradigmatic or paradigmatic-interactional differences in each measurement. For each measure, the results of B1 and B2 are combined. Note that paradigmatic

effects were assessed only in the absence of paradigmatic-interactional effects. For the TBI variable, only data from articulatory sessions were examined.

	Paradigmatic	Paradigmatic-interactional	Either
Articulatory measures	16	9	25/72 (34.7%)
Durations	6	3	9
Amplitudes	6	4	10
Velocities	4	2	6
Acoustic durations	12	16	28/72 (38.9%)
Coda	5	5	10
Vowel	2	4	6
Rime	5	7	12
TBIs	1	7	8/12 (66.7%)
F0	49	15	64/138 (46.4%)
Vowels adjacent to boundaries	41	8	49/92
Changes across boundaries	8	7	15/46

We detected paradigmatic-interactional effects in about 16% of our data (47/294); however, an examination of effect sizes showed that the interaction effects in some of these cases were indeed quite small. For example, see the interaction effect found in the release duration of Participant 1 at B1 and the release amplitude of Participant 2 at B1 in Fig. 10, where paradigmatic and paradigmatic-interactional effects were detected. The difference in the magnitude of the slope between the two regression lines (NRRC vs. RRC) was 0.02 for P01 and 1.1 mm/s for P02. As in these cases, the fact that many interaction effects were quite small is one reason that we do not view them as compelling evidence for categorically distinct prosodic organization. Moreover, the same ambiguity in interpretation of paradigmatic effects applies to paradigmatic-interactional effects: it is not possible to determine whether such effects are due directly to syntactic/semantic differences or whether they are mediated by categorical differences in prosodic organization. Thus, paradigmatic-interactional differences are not considered as robust evidence for distinct prosodic categories.

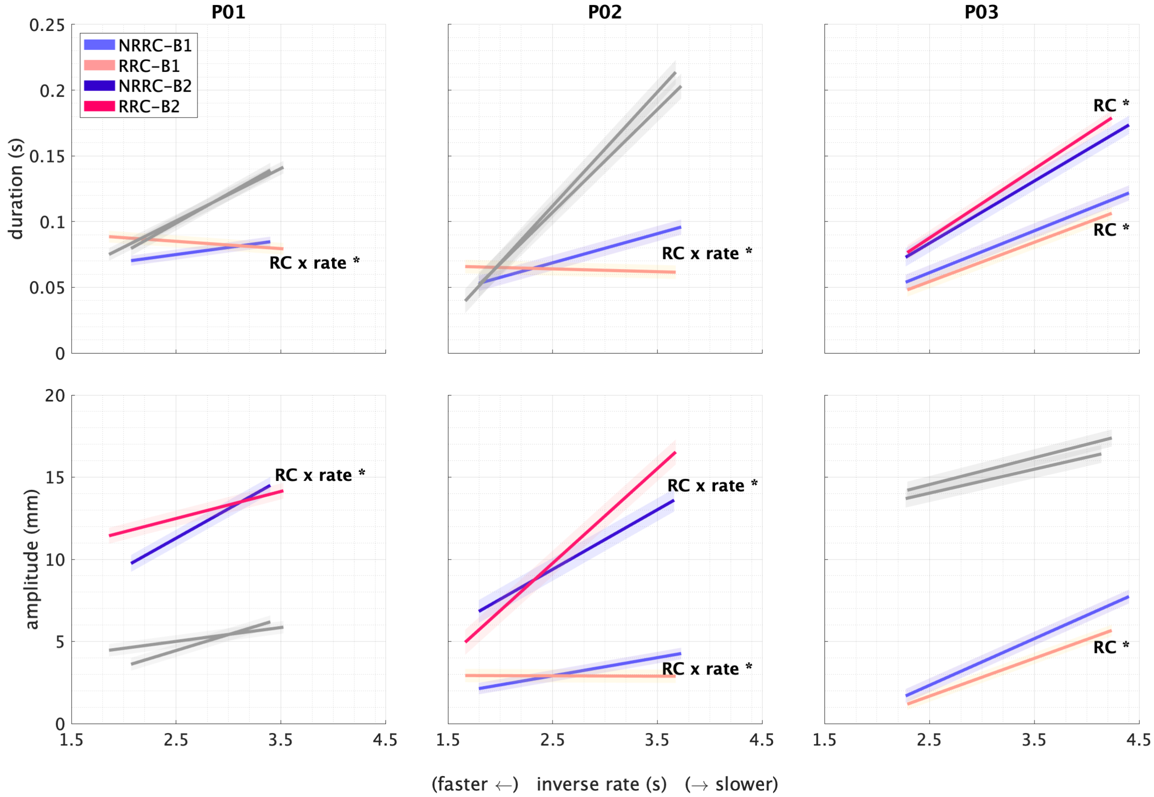

Fig. 10. Paradigmatic and paradigmatic-interactional analyses of articulatory duration and amplitude of the release gesture in Participants 1, 2, and 3. The top figures show the results of the release duration, and the bottom ones show the results of the release amplitude. The colored lines are the linear model fits of the data that showed significant paradigmatic or paradigmatic-interactional differences along with 95% confidence intervals. Specifically, the data from Participants 1 and 2 showed interaction effects (RC × rate *), while the data from Participant 3 showed the effect of RC type (RC *). The gray lines indicate the model fits which did not show statistically significant effects.

4.2. Within-context mixture analyses

Within-context mixture effects were observed in some of our data, but it is unclear whether these occurred frequently enough to provide compelling support for Hyp. 2. A total of 588 cases were examined, and 121 of them (20.6%) showed substantial improvement in the mixture model over the linear one. We examined the prevalence of detected effects by boundary (see Table 4) and found that the measures at B2 are more likely to exhibit mixture effects than B1 ($\chi^2(1, N = 588) = 5.73, p = 0.017$). Out of 121 mixture cases, 71 cases were found at B2, while 50 cases were found at B1. As shown in Table 4, we found such differences between boundaries for most categories of variables.

Table 4. Number of cases that showed significant within-context mixture effects in each measurement at each boundary. For each measure and boundary, the results of NRRC and RRC are combined.

	B1	B2	B1 + B2
Articulatory measures	7/72	18/72	25/144 (17.4%)
Durations	3	5	8
Amplitudes	3	7	10
Velocities	1	6	7
Acoustic durations	14/72	19/72	33/144 (22.9%)
Coda	4	8	12
Vowel	5	3	8
Rime	5	8	13
TBIs	4/12	6/12	10/24 (41.7%)
F0	25/138	28/138	53/276 (19.2%)
Vowels adjacent to boundaries	18	18	36
Changes across boundaries	7	10	17

The proportion of within-context effects was compared across variables (see Table 4), and we found more mixture effects in TBIs than other variables. Note that this is similar to what we found in the paradigmatic and paradigmatic-interactional analyses. More specifically, when within-context effects were examined in each phonetic measure at each boundary, the TBI variable and articulatory movement amplitude of the release gesture at B2 showed within-context effects in 50% of the regressions conducted. These variables were followed by coda and rime durations at B2, TBI at B1, and mean F0 of the vowels at the post-boundary region of B2, which showed mixture effects in more than 30% of the data. However, there were also measures (i.e. the onset-to-release interval at B1 and velocity of the closure gesture at B1) which showed no within-context mixture effects.

An examination of the regression coefficients of mixture models found several qualitatively different patterns. Only cases in which mixture effects were detected are considered here. Specifically, the patterns can be categorized into two groups, according to the relative values of category-specific slope parameters. As shown in Fig. 11, these are slow-rate responsive (1a, 1b) and fast-rate responsive (2a, 2b) patterns. The slow-rate responsive pattern involves model fits where the magnitude of the slope of category S (category associated with slow rate) is larger than category F (associated with fast rate) (i.e. $|\beta_{1F}| < |\beta_{1S}|$), while the fast-rate responsive pattern refers to the opposite case where the magnitude of the slope of category F is larger than category S (i.e. $|\beta_{1F}| > |\beta_{1S}|$). Each category can be further divided into cases where the slope of the more responsive category is positive or negative (e.g. in fast-rate responsive pattern, $\beta_{1F} > 0$ vs. $\beta_{1F} < 0$). See the right panel of Fig. 11 for the model fits that represent each category.

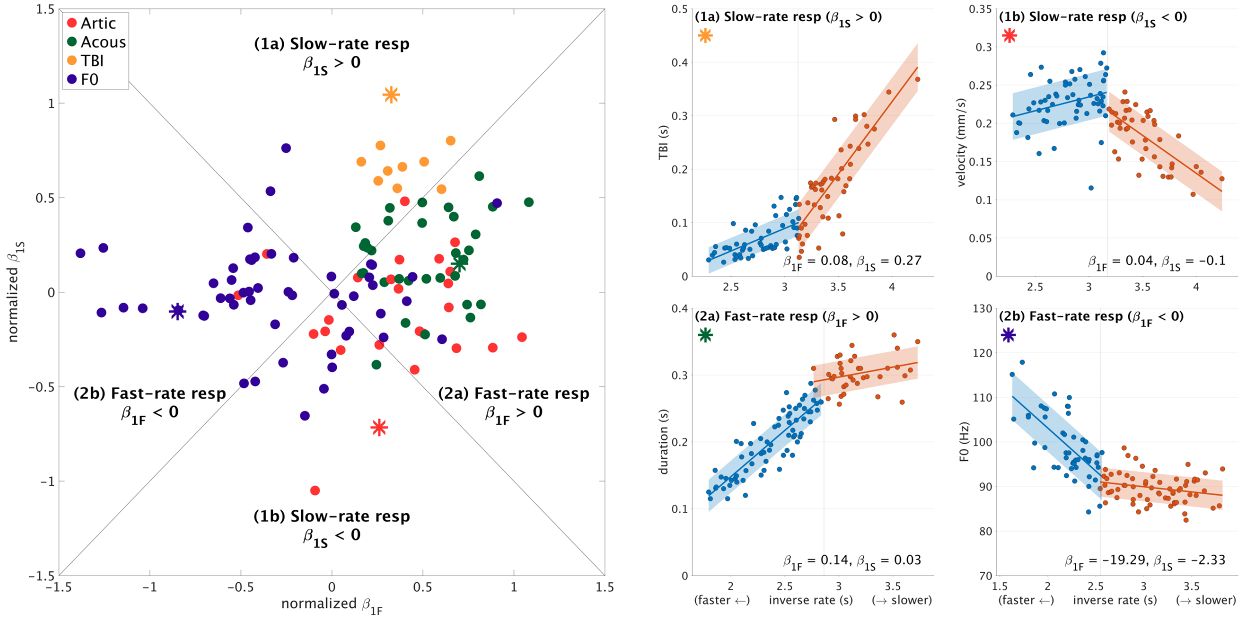

Fig. 11. Patterns of mixture model fits. The left panel shows the relation between the slope of category F and category S in cases where significant within-context mixture effects were detected. Each dot represents an individual case where mixture effects were observed (*n* = 121). In order to compare the slope relations across different phonetic variables, the slopes in this figure are normalized such that the original slopes were divided by the difference between the 5th and 95th percentile of the dependent measures. The panels on the right show examples of each pattern; these are the cases that are marked as asterisks in the left panel. (1a): Participant 3, RRC, TBI at B2; (1b): Participant 3, RRC, movement velocity at the release gesture at B2; (2a): Participant 2, NRRC, rime duration at B1; (2b): Participant 7, NRRC, maximum F0 of the vowel at the pre-boundary region of B1.

Interestingly, phonetic variables differed in what kind of model fits they primarily exhibit. TBIs exclusively showed the pattern of (1a) in Fig. 11 where the slow-rate category is more responsive than the fast-rate category with positive rate effect. These are illustrated in Fig. 12, which shows the mixture model fits of TBI at B2. Except for the NRRC of Participant 6, all of them showed a slow-rate responsive pattern. The mixture model fits of the TBI variable can be interpreted as evidence of distinct prosodic categories, but alternatively, they might arise from a scale-attenuation effect which limits the amount of overlap between gestures at fast rates (i.e. cannot have too much overlap at fast rates). The latter interpretation will be further discussed in Section 5.

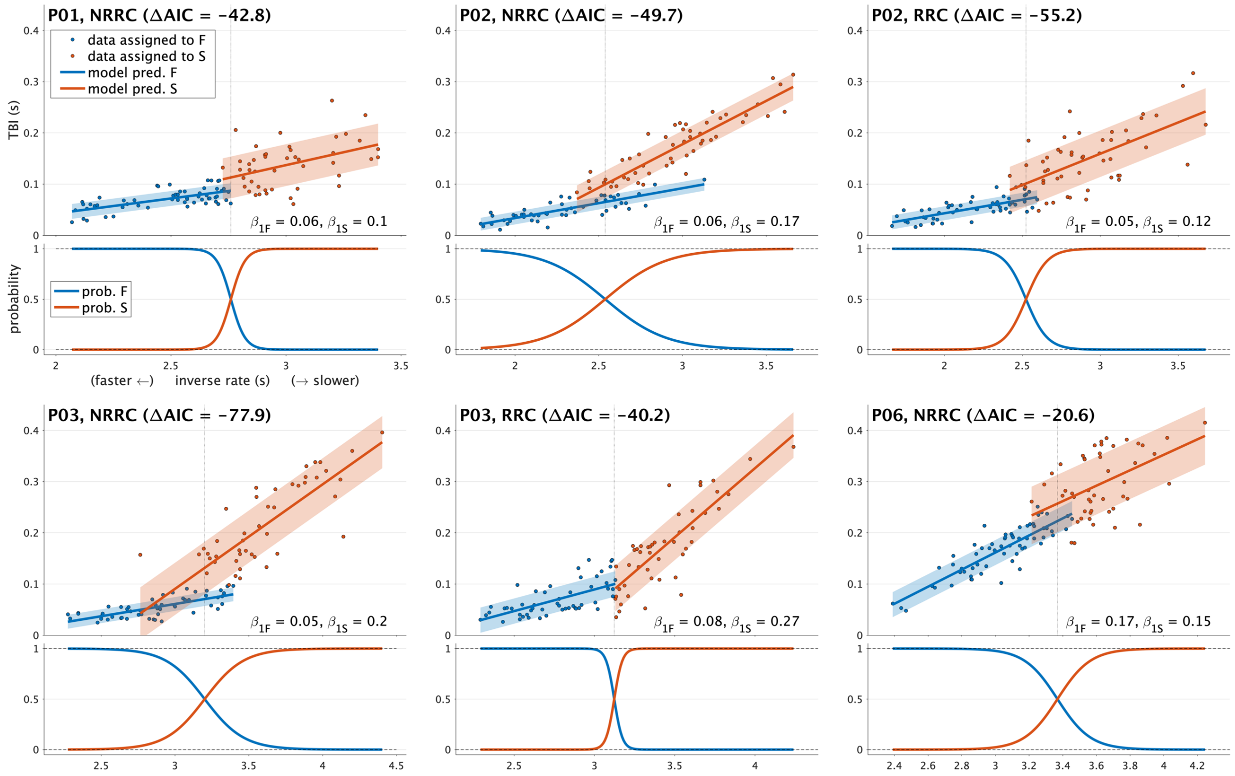

Fig. 12. Mixture model fits of TBI measures at B2. The figure only shows cases that had substantial improvement in the mixture model over the linear one. The upper panels show the actual datapoints and the estimated regression mixtures as well as the slope parameter of each category. The shaded areas are ±1.0 standard deviation. The lower panels show the probability functions. The title of each subfigure shows information about participant and syntactic structure along with the AIC values.

Articulatory variables and acoustic durations showed more varied patterns, but the majority of them showed a fast-rate responsive pattern with positive rate effect. Fig. 13 shows the model fits of the articulatory movement amplitude of the release gesture at B2. Here, the fast-rate responsive pattern with a positive slope (Fig. 11: 2a) was found in Participants 1 (RRC), 2 (RRC), and 3 (NRRC/RRC), but Participants 4 (NRRC) and 6 (NRRC) showed the slow-rate responsive pattern with a negative slope (Fig. 11: 1b). These mixture model fits provide evidence that there exist two distinct prosodic categories, but we cannot rule out the alternative interpretation of scale-attenuation effects. The interpretation of these patterns is complicated by the fact that movement amplitude depends on several factors: the starting position at movement onset, the gestural target, and the gestural duration and overlap with subsequent gestures. In most cases, the patterns suggest a ceiling effect at slow rates, and the fast-rate responsiveness of this variable may be attributable to increases in overlap or target undershoot at fast rates, or to differences in LA or TT aperture at movement offset, at fast vs. slow rates.

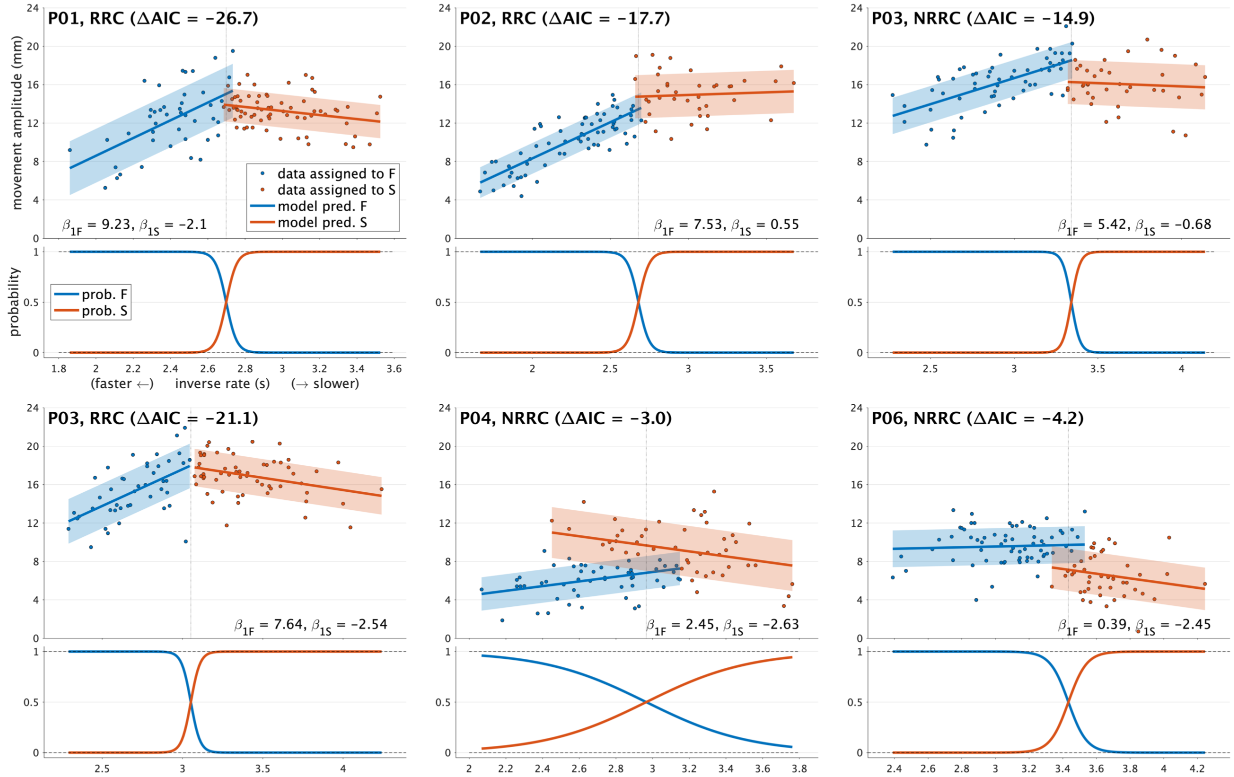

Fig. 13. Mixture model fits of the articulatory movement amplitudes of the release gesture at B2.

Lastly, in mixture model fits of F0 measurements, the majority of cases exhibited a fast-rate responsive pattern with negative rate effect (Fig. 11: 2b). Fig. 14 shows the estimated regression mixtures for maximum F0 of the vowels that occurred at the pre-boundary region of B1 (i.e. /a/ of the first target word). We observed the fast-rate responsive, negative slope pattern of (2b) in Fig. 11 in both syntactic contexts of Participant 7, RRC of Participant 10, and NRRC of Participant 12. However, we also identified the slow-rate responsive pattern with negative rate effect (Fig. 11: 1b) in the RRC of Participant 1 and positive rate effect (Fig. 11: 1a) in the NRRC of Participant 10. The mixture model fits of this measure may be interpreted as evidence of distinct prosodic categories with a category-specific rate effect. However, another possible explanation is that as rate increases, there is more time for the L pitch accent gesture in a phrase-final L* or H*+L gesture to reach its target: this can account for why the scale appears to attenuate at slower rates.

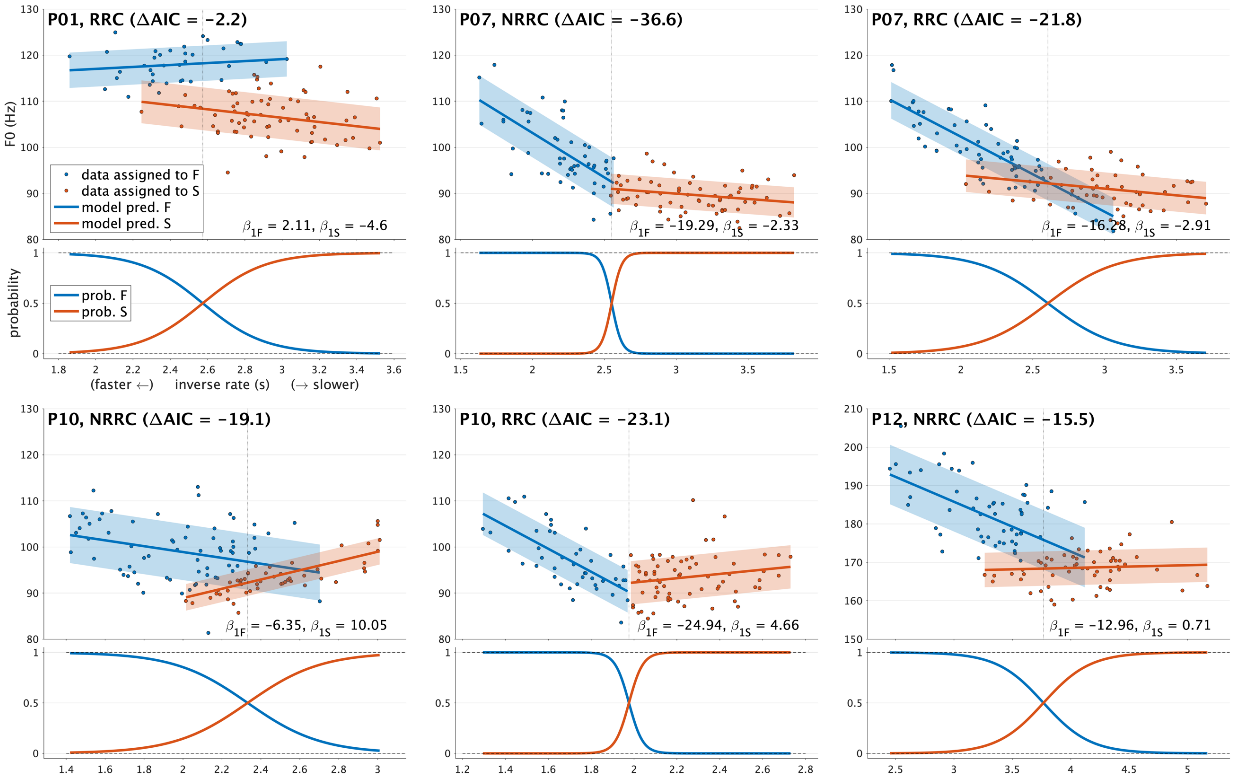

Fig. 14. Mixture model fits of the maximum F0 of the vowels at the pre-boundary region of B1.

In addition to the analyses of within-context effects by boundary and by variable, we also conducted analyses by participant. No participants showed within-context effects in more than 50% of the data. Individual differences were observed such that some participants (Participants 2, 7, 8) showed a clear separation of categories in at least 30% of cases, while other participants (Participants 4, 5, 11) showed two categories in less than 10% of cases (see dark blue bars in Fig. 15).

4.3. Analyses with inverse rate vs. proper rate

Analyses conducted with proper rate as an independent variable showed overall similar results to the analyses with inverse rate presented in Sections 4.1 and 4.2; however, we found some differences in the pattern of the mixture model fits in acoustic variables. First, the results of the paradigmatic and paradigmatic-interactional analyses were very similar such that 96.3% of the cases (283/294) showed matching results. The TBI showed the highest proportion of paradigmatic and paradigmatic-interactional differences among the phonetic variables in proper rate analyses, as was the case for in inverse rate analyses presented in Section 4.1.

With respect to the within-context effects, 89.1% of the cases (524/588) showed matching results such that when a within-context effect was detected in inverse rate analyses, it was also detected in proper rate analyses and vice versa. There were 39 cases (6.6%) of non-matching results where the within-context effect was found only in the inverse rate analyses, and 25 cases (4.3%) where the effect was found only in the proper rate analyses. There was no apparent pattern in the distribution of these non-matching cases across participants or variables.

Within-context effects examined by boundary, by variable, and by participant were similar between inverse rate and proper rate analyses. In proper rate analyses, the mixture effects were more likely to be observed at B2 than at B1 ($\chi^2(1, N = 588) = 6.14, p = 0.013$), as was the case for inverse rate analyses. Likewise, TBI showed the highest proportion of within-context effects among all variables in the proper rate analyses. In particular, TBI, movement amplitude of the release gesture, and coda duration at B2 showed within-context effects in more than 30% of cases in both inverse rate and proper rate analyses. Among these three variables, the TBI and movement amplitude showed stronger within-context effects (more than 40% of cases). To some extent, participants also showed different proportions of mixture effects in proper rate analyses; see the pink bars in Fig. 15. Fig. 15 also compares the proportion of data that showed significant within-context effects between the two rate analyses within each participant. Note that for some participants – Participants 7, 9, or 11 – we observe a large difference in proportion of within-context effect depending on whether the analyses were conducted with inverse rate or proper rate as an independent variable.

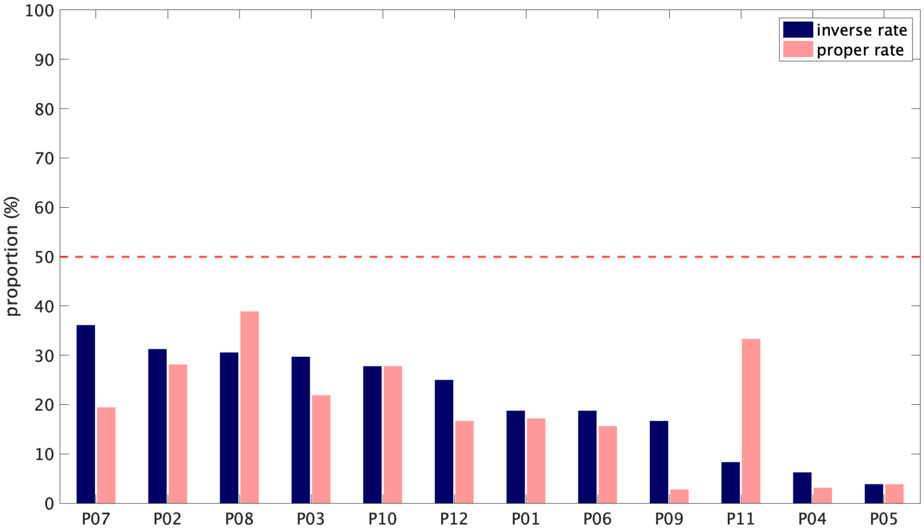

Fig. 15. Comparison of the proportion of within-context effects between inverse rate and proper rate analyses within each participant. The data are sorted in the order of the proportion of within-context effects found in inverse rate analyses (dark blue bars). The red dashed line marks 50% of the data.

Comparing the category-specific regression slopes of inverse rate and proper rate analyses, we found differences particularly in acoustic durations. Specifically, while the majority of the acoustic durations exhibited the *fast*-rate responsive pattern (Fig. 11: 2a) in inverse rate analyses, they showed the *slow*-rate responsive pattern (Fig. 11: 1a) in proper rate analyses. See Fig. 16 which compares the mixture model fits of rime durations at B1 across the two rate analyses. In inverse rate analyses, out of 33 acoustic duration measures that showed mixture effects, 25 cases showed the fast-rate responsive pattern, while eight cases showed the slow-rate responsive pattern (see green dots in Fig. 11). However, in proper rate analyses, mixtures were detected in 35 cases, and 25 of them showed the slow-rate responsive pattern, while 10 cases showed the fast-rate responsive pattern. Besides acoustic durations, other variables showed similar patterns of mixture model fits between the two rate measures.

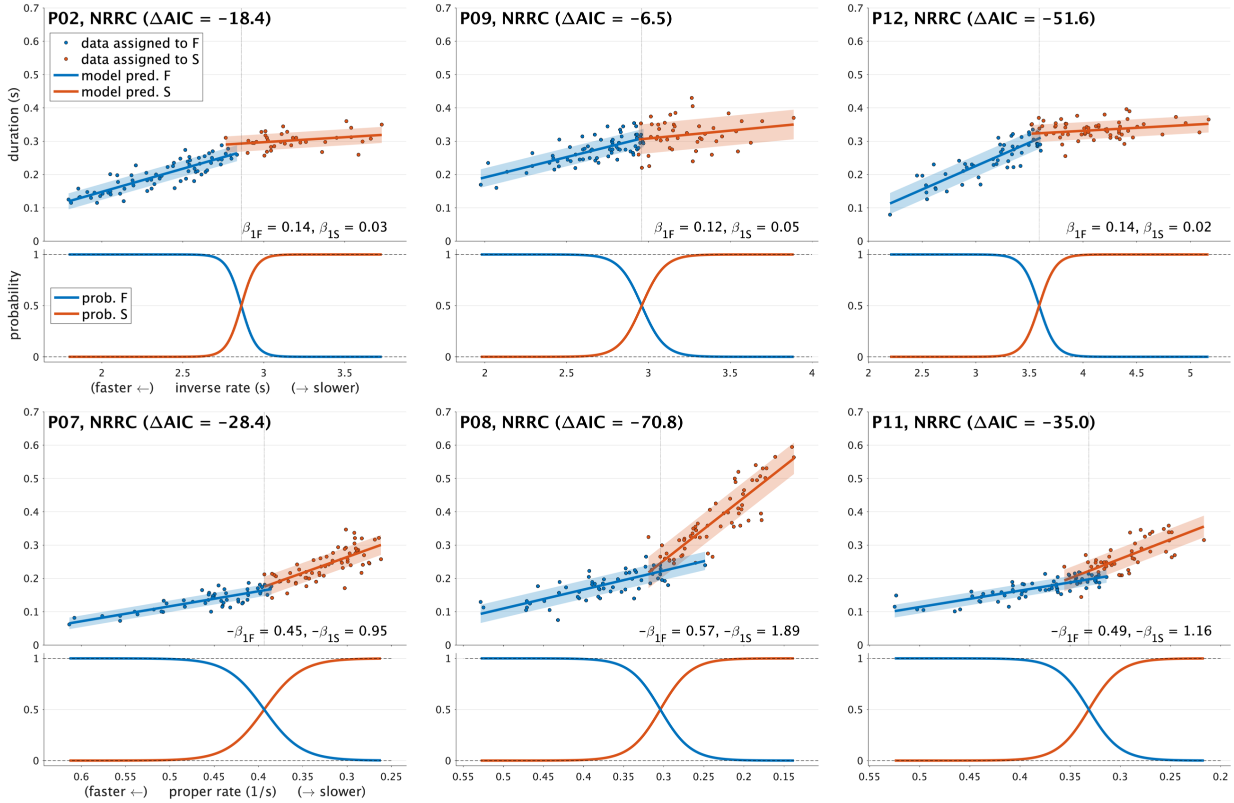

Fig. 16. Comparisons of the mixture model fits of rime durations at B1. The top figures are the results of the analyses conducted with inverse rate, and the bottom figures are from the analyses conducted with proper rate. For an easier comparison between the analyses of the two rate measures, the orientation of proper rate coordinates is reversed, and we examined $-\beta_{1F}$ and $-\beta_{1S}$ in the proper rate analyses.

5. Discussion

Overall, we did not observe strong phonetic evidence for hierarchical organization of prosodic phrases. The most robust evidence was observed in TBIs, where 41.7% (10/24) of the participants/contexts/boundaries exhibited mixture effects in inverse rate analyses. The investigation of within-context effects in each phonetic measure at each boundary found a relatively high proportion of mixture effect at TBI, movement amplitude of the release gesture, and coda duration at B2. One feature that unifies these measures is that they reflect articulatory timing and overlap at phrase boundaries. For other measures, evidence for categorical variation in prosodic organization was more sporadic. Furthermore, we found that the choice of independent variable—inverse rate (sentence duration) or proper rate (sentences/s)—had consequences for the patterns of the mixture model fits. Below we elaborate on these findings.

First, the paradigmatic and paradigmatic-interactional comparisons showed a moderate number of cases in which significant paradigmatic or paradigmatic-interactional effects were detected. In the inverse rate analyses, 42.5% of cases showed significant effects. However, for the reasons that we discussed in Section 2.2, the presence of paradigmatic and paradigmatic-interactional effects cannot in itself be taken as evidence of hierarchical phrase organization, because these effects could alternatively arise directly from syntactic, semantic, or informational factors which modulate articulation at the boundaries of a non-hierarchical phrase structure. In other words, the interpretation of such effects as evidence for hierarchical phrase structure requires an unsubstantiated assumption that distinct prosodic phrase structures are responsible for the effects. Furthermore, we noticed that some of the interactional effects were not very substantial, which calls into question the extent to which inferences about category-specific differences in rate effects should be drawn from these analyses.

Second, the within-context mixture analyses provided some evidence for distinct prosodic categories, but this evidence was not consistent across participants, measures, or boundaries. Regarding boundaries, we found that there were more cases in which rate-dependent mixtures were identified for variables at B2 than B1 (B1: 41.3% vs. B2: 58.7%). This could be interpreted to mean that rate-conditioned differences in prosodic organization are more likely to be observed at B2 than B1. This could arise if, for example, a higher-level phrase boundary was adopted at B2 for slower rates, while a lower-level boundary was adopted for faster rates. This would be the case if the sentence is organized as a single IP for fast rates but as two IPs for slow rates, as (1) below:

(1)

fast rates: **[[A Mr. Hodd,** $_{ip}$**]** **[who knows Mr. Robb,** $_{ip}$**]** **[often plays tennis.** $_{ip}$**]** $_{IP}$**]**
slow rates: **[[A Mr. Hodd,** $_{ip}$**]** **[who knows Mr. Robb,** $_{ip}$**]** $_{IP}$**]** **[[often plays tennis.** $_{ip}$**]** $_{IP}$**]**

Regarding differences between participants, we saw that there were several participants who exhibited a larger proportion of within-context effects than others; for instance, Participants 2, 7, and 8 showed mixture effects in more than 30% of the data in inverse rate analyses. These differences could be interpreted to mean that those participants were more prone to using different prosodic organizations as a function of speech rate, along the lines illustrated in (1). The fact that there were differences between participants was not surprising, since we anticipated that participants would differ in the extent to which variation in speech rate influences their prosodic organizations.

One possible source of between-participant differences is that participants may be more likely to employ different prosodic phrase structures (and thus exhibit more within-context effects) if they have a wider range of empirical speech rate. Thus, we conducted a post-hoc analysis of whether by-participant proportions of mixture effects are related to range of speech rate. This was accomplished by conducting a linear regression between by-participant rate ranges and proportions of mixture effects. The results showed that participants' range of inverse speech rate was not a significant factor in predicting the proportion of within-context effects in their data. However, this hypothesis was supported when the proper rate measure was used as an independent variable ($F(1,10) = 7.97, p < 0.05, R^2 = .44$).

Regarding differences between variables, we found that TBI, movement amplitude of the release gesture, and coda duration especially at B2 were the variables that most often exhibited evidence of within-context rate-dependent mixtures in both inverse rate and proper rate analyses. This could be interpreted to indicate that these variables are more strongly modulated by differences in prosodic organization. It is important to note that all three of these variables reflect boundary-localized changes in articulatory timing and/or gestural overlap, and thus the findings suggest that such measures may be the most fruitful ones to investigate in future studies. Other variables showed weaker evidence of mixtures in either analysis.

In general, we ask why evidence for mixtures was not more robustly present across participants, variables, or boundaries. One possible interpretation is that categorical differences in prosodic organization simply do not exist — boundary modulations may be driven by other factors such as syntactic structure, semantic difference, or information status, which modulate the phonetic parameters associated with non-hierarchical phrase structure (see Section 2.2). In this interpretation, the mixtures that were identified must arise from alternative mechanisms, since these factors do not vary within a given syntactic context. One possible mechanism that we discuss below is scale attenuation: there may be upper or lower bounds on variables that derive from control mechanisms rather than a hierarchy of phrase categories.

A different explanation for the moderate rate of mixture detection is that the statistical power of the models is not high enough to detect mixtures in all cases. In general, each of the within-context analyses employed about 120 observations. This might not be enough to identify mixtures if the effect sizes are relatively small, and indeed, our simulations reported in Section 2.3 (and Appendix: Mixture model details) suggest that this is the case. Future studies may benefit from using the effect size estimates obtained in this study to conduct numerical simulations to determine statistical power.

Another possible explanation for the absence of more robust mixtures is that the range of rates elicited for each participant/context was not wide enough to obtain a substantial proportion of responses in each category of the mixture. We imposed the external criterion that at least 33% of observations had to be associated with each category. We believe this criterion is reasonable given our methodological efforts to elicit a wide range of variation in speech rates. Indeed, we saw that the average range of inverse rate across participants was 2.41 s, with the minimum of 1.67 s and the maximum of 5.36 s. This makes us fairly confident that all participants produced a range of rates such that the fastest and slowest rates extended beyond those that are typically observed in conversational speech. Thus, we do not think this alternative explanation is very plausible.

In examining the category-specific rate effects present in mixture models, we identified two main patterns: a "slow-rate responsive" pattern in which speech rate had a stronger influence on the slow-rate category and a "fast-rate responsive" pattern in which the rate effect was stronger for the fast-rate category. These patterns can also be characterized according to the slopes of the rate effect for the responsive category. This distinction is evident in Fig. 11 where the category-specific rate slopes of each analysis were plotted against one another. Specifically, we found that the TBI has a prevalence of slow-rate responsive patterns where the TBI increases more substantially at slow rates. In contrast, acoustic durations and articulatory variables showed fast-rate responsive patterns with a positive rate slope. The majority of F0 cases showed fast-rate responsive patterns with a negative slope. The co-existence of such patterns is interesting: what is its origin?

One important consideration in interpreting the mixture model slopes and category responsiveness is that a significant mixture model fit can arise not only from the presence of distinct prosodic categories, but also arise from floor and/or ceiling effects—i.e. scale attenuation—which do not necessarily entail two distinct categories. For example, it may be that there are physiological mechanisms or cognitive constraints (e.g. limits on target values or gestural overlap) which prevent certain variables from having values that are too large or small (e.g. upper bounds on articulator movement speeds, or lower bounds on F0 for modal voicing). Such effects could arise from hard limits on values, or more likely from attenuation effects which grow stronger as distance beyond a threshold increases. Fig. 17 shows how scale attenuation patterns can be mis-identified as mixtures. The top figures of Fig. 17 illustrate data in which artificial floor and/or ceiling effects were imposed on a variable. Floor effects were accomplished by generating random variables that were linearly related to speech rate and then increasing values which were below a threshold, with the increase being a random value from 50% to 150% of the distance beyond the threshold. Ceiling effects were generated similarly for values above a threshold. Spline fits of the data illustrate the resulting scale attenuation. In all three cases, the regression mixture model is a substantial improvement over a linear model, even though two categories were not explicitly present in the model which generated the data. It is not possible with our method to distinguish the scale attenuation effects of (A, B, and C) from a rate-conditioned difference in prosodic organization.

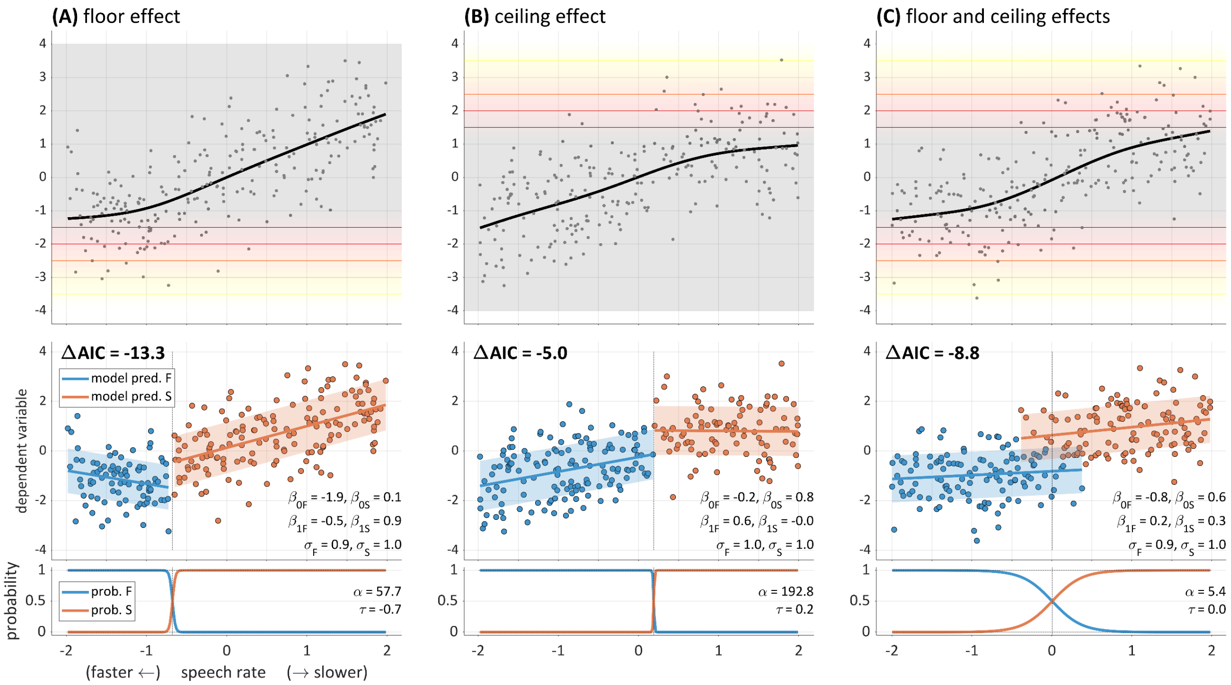

Fig. 17. Illustration of floor and ceiling effects. Top figures show spline fits of scale-attenuated data and the scale-attenuating force fields. Bottom figures show mixture regressions of the scale-attenuated data (upper panels) and category probability functions (lower panels). Simulations were conducted with Gaussian-distributed dependent and independent variables. (A) floor effect; (B) ceiling effect; (C) combined floor and ceiling effects.

The potential conflation of prosodic category mixtures with scale attenuation entails that even the strongest form of evidence for categories—within-category rate-conditioned mixture effects—is not unambiguous. However, scale attenuation may be ruled out in some cases if it can be shown that the range of values of a distribution is exceeded by a given participant in some other contexts. For instance, in considering F0 measures, when the range of values for a given measure is exceeded for a similar measure in a different context, scale attenuation can likely be ruled out. General procedures to identify scale-attenuation are highly desirable.

In this context, the co-existence of slow- and fast-rate responsive mixtures may simply reflect different distributions of categories associated with various phonetic variables, but also may be interpreted as different forms of scale attenuation. For the fast-rate responsive patterns, where rate has a smaller effect for the slow-rate category, these could indicate scale-attenuation at slow rates. Conversely, the slow-rate responsive patterns, which have smaller rate effects for fast rates, could arise from scale attenuation at fast rates. More specifically, the TBI had a prevalence of slow-rate responsive mixtures. This suggests that there are constraints which impose a maximum allowable overlap between gestures across a boundary. At fast rates, the overlap measure attenuates by reaching a floor. On the other hand, most acoustic and articulatory variables as well as F0 had a prevalence of fast-rate responsive patterns. We can interpret these as cases in which tract variables or F0 may be more likely to reach target values at slow rates and fail to do so at fast rates, due to target undershoot or gestural overlap.

From theoretical perspective, the fact that the within-context effects are not robustly present in the data and may be due to scale attenuation is not necessarily surprising. Models of articulatory control are generally agnostic regarding underlying categorical differences in prosodic structure, and hence do not inherently favor hierarchical or non-hierarchical conceptions of phrase structure. The articulatory models that are most relevant to the interpretation of boundary-related prosodic variation are π-gesture model of Articulatory Phonology as developed in Byrd and Saltzman (2003) and the attentional modulation model of selection-coordination theory described in Tilsen (2018). Although the mechanisms that lead to boundary-related effects are different in the two models, they both have continuous parameters which regulate such effects. In the case of π-gesture model, these parameters are the activation levels and temporal extents of π-gestures; in the attentional modulation view, the parameter is the relative reliance on external vs. internal sensory feedback and the values of feedback thresholds required for suppression of active gestures. Whether or not there exists a hierarchy of phrase types is thus logically independent from the models, since the continuous parameters of the models could exhibit categorical distributions associated with prosodic phrase categories or gradient variation due to other factors.

Speech rate plays an important role in our methods for detecting evidence for categorical differences in prosodic organization. We examined whether different syntactic/semantic contexts interact with speech rate in different ways (i.e. paradigmatic-interactional comparison) and also examined whether there is a rate-dependent mapping of a given syntactic context to different prosodic organizations. For these analyses, eliciting a continuous variation of speech rate was crucial. By using a novel experimental design which involved a moving visual analog rate cue, we were able to elicit a wide variation in speech rate within each participant. This method better reflects the continuous nature of speech rate variation, and it allowed us to mitigate issues that arise from participant-specific interpretations of categorical rate instructions.

We also compared the results of analyses conducted with inverse rate and proper rate as independent variables and found that the two types of analyses showed both similarities and differences. The results of the paradigmatic and paradigmatic-interactional analyses were very similar such that the results were same in 96.3% of the cases (283/294). Within-context analyses also showed high similarity as 89.1% of the cases (524/588) showed matching results. In both analyses, we found more mixtures at B2 than at B1 and the highest proportion of mixture effects in the TBI, movement amplitude of the release gesture, and coda duration at B2. Furthermore, we found differences in the proportion of within-context effects across participants in both analyses. The most important difference that we identified was in the relative slopes of regression coefficients for acoustic durations. While we found a fast-rate responsive pattern in the majority of acoustic duration measures in inverse rate analyses, we found a slow-rate responsive pattern in the proper rate analyses. It is interesting that such different patterns are observed within the same variable, and moreover, the difference is found only in acoustic durations. We suggest that future studies should carefully consider how speech rate measures are defined and assess whether alternative definitions influence results.

6. Conclusion

While many previous studies have presupposed that hierarchical prosodic phrase structure exists, this study took a step back from this assumption and instead examined phonetic evidence for categorical variation in phrase types from a more theory-neutral perspective. We proposed that there is a range of the quality of evidence for the existence of hierarchical prosodic phrase structure: paradigmatic differences are relatively weak evidence, paradigmatic-interactional differences are somewhat stronger evidence, and within-context rate-conditioned mixture effects are the strongest form of evidence. Although the results found a fair amount of paradigmatic and paradigmatic-interactional differences, due to ambiguity in how these effects are interpreted, they should not be taken as strong evidence for prosodic categories. Within-context effects were also found in some of our data, yet because these effects were relatively sporadic and may alternatively arise from scale attenuation, they provide moderate but not entirely convincing evidence for categorically distinct prosodic organization. The findings from our analyses thus suggest that there is no indisputable form of evidence that supports the presence of hierarchical phrase categories. The absence of more robust evidence of categorically distinct prosodic organizations calls into question the nature of prosodic structure – are prosodic phrases really organized in a hierarchical way, or do differences in variables arise from gradient modulations of parameters at the boundaries of non-hierarchically organized phrases? The findings challenge the common assumption that phonetic variables associated with phrase boundaries reflect categorical differences in phrase types.

One possible direction that is worth investigating in the future is whether we can find evidence of hierarchical phrase organization in higher-dimensional representations of the information at or near prosodic boundaries. It may be the case that the lack of robust evidence for phrase categories is due to our focus on individual measures. Although cues for phrasal organization may be redundant, such cues may add up to mark prosodic boundaries. Thus, it would be interesting to examine whether combinations of different cues show stronger evidence of hierarchical phrase structure. In that sense, one might examine how listeners perceive boundary strengths in the data we collected, in order to assess whether there is a more holistic difference in prosodic organization conditioned on speech rate. Another possibility is to use machine learning or neural networks with high-dimensional input to infer whether there are multiple categories of phrase types.

Other important contributions of this study include the use of an experimental design that elicits continuous variation of speech rate, and a comparison of analyses conducted with inverse rate and proper rate measures. We found that there were indeed some cases in which the choice of rate measure affects the outcome of analyses, especially in terms of the mixture model fits. We cannot readily account for why these differences occur and also why the difference is found primarily in acoustic durations; nonetheless, the results highlight the importance of the choice of rate measure in analyses. Further investigations are needed to better understand how participants control their rate of speech and which rate measure is best-suited for measuring that control.

Acknowledgements

We would like to thank Draga Zec, Marten van Schijndel, and members of the Cornell Phonetics Lab for discussions of this research. We also thank Bruce McKee for technical support in conducting the experiments.

Appendix: Mixture model details

To assess whether the mixture models correctly identify mixtures in simulated data, we conducted regressions for simulations that varied the following: α_1, the steepness of the category probability function; Δμ, the difference in category intercepts (i.e. $\beta_{0B} - \beta_{0A}$); and the within-category slope, which was constrained to be the same for both categories (i.e. $\beta_{1A} = \beta_{1B}$). One-hundred simulations were conducted for each combination of parameters listed in Table A1. In order to be counted as evidence for two categories, the criteria described in Section 3.3 had to obtain. The proportions of cases which met the criteria are shown for each combination of parameters. Observe that in all cases where the category intercepts did not differ ($\Delta\mu = 0$), evidence for a mixture was not detected. Conversely, when the category intercepts differed substantially ($\Delta\mu = 2$) and were well separated ($\alpha_1 = 10.0$), the mixture detection had a high success rate. This success rate depended on the within-category rate effect slope.

Table A1. Results for validation of mixture models

α_1	$\Delta\mu$	β_1	mixture detected	model converged	ΔAIC criterion	transition location criterion	cluster size criterion	extremal probabilities criterion
0	0	0	0.02	0.93	0.11	0.92	0.53	0.8
0	0	1	0.05	0.97	0.14	0.95	0.54	0.82
0	0	2	0.01	0.97	0.13	0.9	0.46	0.8
0	1	0	0.04	0.96	0.14	0.98	0.53	0.9
0	1	1	0.07	0.95	0.17	0.94	0.48	0.89
0	1	2	0.06	0.96	0.12	0.94	0.55	0.87
0	2	0	0.03	0.97	0.12	0.94	0.6	0.83
0	2	1	0.04	0.96	0.15	0.92	0.66	0.84
0	2	2	0.03	0.97	0.15	0.93	0.5	0.83
1	0	0	0.02	0.93	0.09	0.92	0.55	0.86
1	0	1	0.01	0.99	0.2	0.98	0.51	0.77
1	0	2	0.02	0.95	0.14	0.96	0.52	0.9
1	1	0	0.08	0.97	0.16	0.94	0.53	0.82
1	1	1	0.02	0.97	0.12	0.96	0.53	0.77
1	1	2	0.05	0.99	0.1	0.96	0.53	0.82
1	2	0	0.17	0.99	0.32	0.98	0.55	0.85
1	2	1	0.09	0.95	0.21	0.95	0.6	0.82
1	2	2	0.08	0.99	0.22	1	0.68	0.81
10	0	0	0.05	0.96	0.15	0.97	0.6	0.88
10	0	1	0.03	0.99	0.11	0.92	0.47	0.83
10	0	2	0.05	0.98	0.11	0.93	0.53	0.82
10	1	0	0.34	0.96	0.42	0.99	0.8	0.95
10	1	1	0.35	0.96	0.44	0.97	0.76	0.96
10	1	2	0.4	0.98	0.51	0.99	0.82	0.93
10	2	0	0.99	1	0.99	1	1	1
10	2	1	0.99	0.99	1	1	0.99	1
10	2	2	0.97	0.99	0.99	1	0.98	0.99

Several examples of circumstances in which mixture models may fail to identify evidence for two underlying categories are shown in Fig. A1. Panel (i) shows a case where the category intercepts are not sufficiently different. The corresponding mixture model in (i’) is not a substantial improvement over a simple linear regression model (ΔAIC = 7.9). Panel (ii) shows a case in which the categories are not sufficiently separated according to the independent variable. Once again, the mixture model in (ii’) is not a substantial improvement. Panel (iii) shows a case in which the effect of speech rate obscures the mixture. Here the mixture model was supported vis-a-vis the ΔAIC criterion, but only 16.7% of the datapoints were assigned to category A. We impose an additional criterion that at least 33% of the datapoints should be assigned to each category (see Section 3.3), and hence this is also a case in which the model fails to detect evidence for two categories.

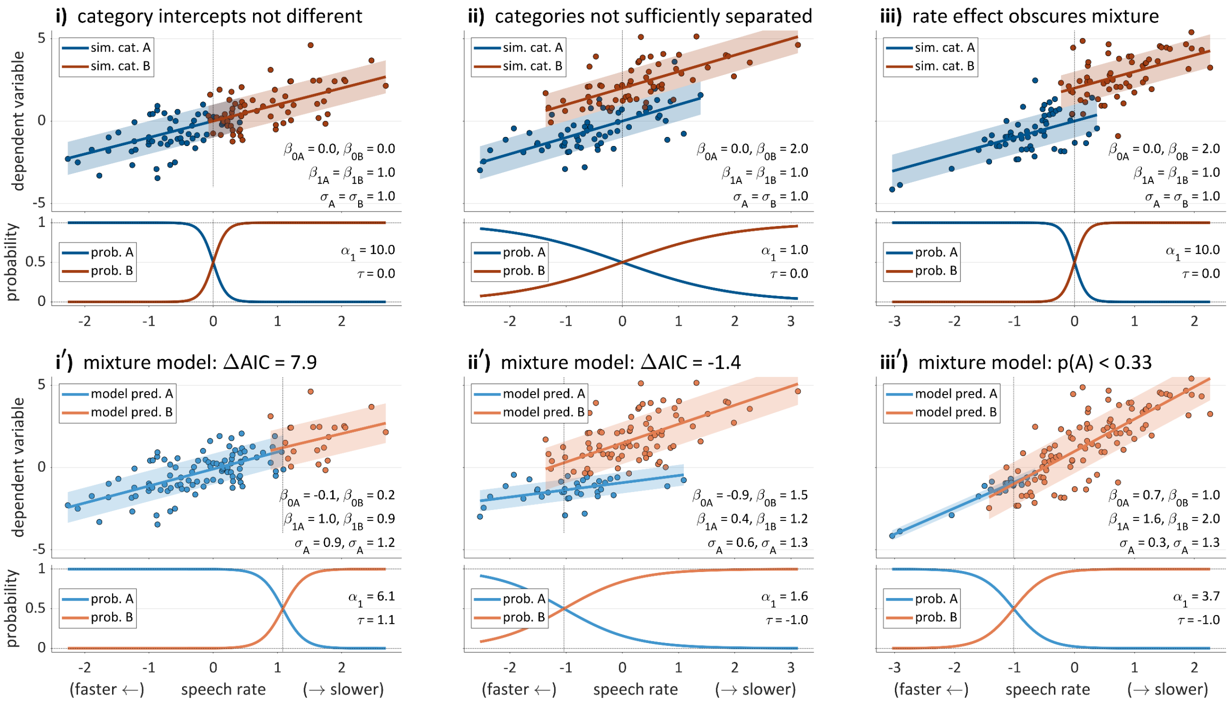

Fig. A1. Examples of failures to identify evidence for two categories. (i, ii, iii): simulated datapoints and parameters (upper panels) along with simulation probability functions and parameters (lower panels). (i’, ii’, iii’): estimated regression mixtures and probability functions with their parameters. (i) and (i’) show the case where the difference of the category intercepts is small, while (ii) and (ii’) show the case where the categories are not well separated. (iii) and (iii’) violate the additional criterion where the smaller category should have at least a third of the datapoints.

References

Arnold, D. (2007). Non-restrictive relatives are not orphans. *Journal of Linguistics*, *43*(2), 271–309.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Berkovits, R. (1993a). Progressive utterance-Final lengthening in syllables with final fricatives. *Language and Speech*, *36*(1), 89–98.

Berkovits, R. (1993b). Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics*, *21*(4), 479–489.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, *57*, 3–16.

Byrd, D., Krivokapić, J., & Sungbok, L. (2006). How far, how long: On the temporal scope of prosodic boundary effects. *The Journal of the Acoustical Society of America*, *120*, 1589–1599.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, *26*(2), 173–199.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180.

Cho, T. (2006). Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English. In L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory Phonology 8* (pp. 519–548). Berlin: Mouton de Gruyter.

Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, *29*(2), 155–190.

Choi, J. (2003). Pause length and speech rate as durational cues for prosody markers. *The Journal of the Acoustical Society of America*, *114*(4), 2395–2395.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper & Row.

Del Gobbo, F. (2007). On the syntax and semantics of appositive relative clauses. In N. Dehé & Y. Kavalova (Eds.), *Parentheticals*. Amsterdam: John Benjamins Publishing.

Demirdache, H. (1991). *Resumptive chains in restrictive relatives, appositives, and dislocation structures* [PhD Thesis]. Massachusetts Institute of Technology.

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *The Journal of the Acoustical Society of America*, *89*(1), 369–382.

Fabb, N. (1990). The difference between English restrictive and nonrestrictive relative clauses. *Journal of Linguistics*, *26*(1), 57–77.

Fery, C., & Truckenbrodt, H. (2005). Sisterhood and tonal scaling. *Studia Linguistica*, *59*(2–3), 223–243.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740.

Grün, B., & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, *51*(11), 5247–5252.

Grün, B., & Leisch, F. (2008). FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, *28*(4), 1–35.

Hayes, B. (1989). The prosodic hierarchy in meter. In P. Kiparsky & G. Youmans (Eds.), *Rhythm and meter* (pp. 201–260). Orlando, FL: Academic Press.

Horne, M., Strangert, E., & Heldner, M. (1995). Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. *In Proceedings of ICPhS*, *95*, 170–173.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*(1), 79–87.

Jun, S.-A. (2005). Prosodic typology. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 430–458). Oxford: Oxford University Press.

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2004). Domain-initial articulatory strengthening in four languages. *Phonetic Interpretation: Papers in Laboratory Phonology VI*, 143–161.

Kim, M., Vermunt, J., Bakk, Z., Jaki, T., & Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 601–614.

Ladd, D. R. (1986). Intonational phrasing: The case for recursive prosodic structure. *Phonology Yearbook*, *3*, 311–340.

Ladd, D. R. (1988). Declination '"reset"' and the hierarchical organization of utterances. *The Journal of the Acoustical Society of America*, *84*(2), 530–544.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent glass regression in R. *Journal of Statistical Software*, *11*(8), 1–18.

McCawley, J. D. (1968). *The phonological component of a grammar of Japanese*. The Hague: Mouton.

McCawley, J. D. (1981). The syntax and semantics of English relative clauses. *Lingua*, *53*, 99–149.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). *The Kaldi speech recognition toolkit*. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, *90*(6), 16.

Roettger, T. B., Winter, B., & Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics*, *73*, 1–7.

Selkirk, E. (1972). *The phrase phonology of English and French.* [PhD Thesis]. Massachusetts Institute of Technology.

Selkirk, E. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry*, *11*(3), 563–605.

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

Selkirk, E. (2005). Comments on intonational phrasing in English. In S. Frota, M. Vigário, & M. J. Freitas (Eds.), *Prosodies: With special reference to Iberian languages* (pp. 11–58).

Syrdal, A. K., & McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. *Sixth International Conference on Spoken Language Processing*.

Tabain, M. (2003). Effects of prosodic boundary on /aC/ sequences: Articulatory results. *The Journal of the Acoustical Society of America*, *113*(5), 2834–2849.

Tilsen, S. (2018). Three mechanisms for modeling articulation: Selection, coordination, and intention. *Cornell Working Papers in Phonetics and Phonology 2018*.

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249–267.

Truckenbrodt, H. (2002). Upstep and embedded register levels. *Phonology*, *19*(1), 77–120.

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445–472.

van den Berg, R., Gussenhoven, C., & Rietveld, T. (1992). Downstep in Dutch: Implications for a model. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in laboratory phonology II: Gesture, segment, prosody* (pp. 335–359). Cambridge: Cambridge University Press.

Wagner, M. (2005). *Prosody and recursion* [PhD Thesis]. Massachusetts Institute of Technology.

Wightman, C. W. (2002). ToBI or not ToBI? *Speech Prosody 2002, International Conference*.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707–1717.