Journal of Phonetics 104 (2024) 101310

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Intonational categories and continua in American English rising nuclear tunes

Jeremy Steffman [a,*], Jennifer Cole [b], Stefanie Shattuck-Hufnagel [c]

[a] The University of Edinburgh, United Kingdom
[b] Northwestern University, Unites States
[c] Massachusetts Institute of Technology, United States

ARTICLE INFO

Article history:
Received 8 March 2023
Received in revised form 6 February 2024
Accepted 15 February 2024

Keywords:
Intonation
Intonational phonology
F0 modeling
Speech perception
Speech production

ABSTRACT

The present study tests a prediction from the prevalent Autosegmental-Metrical (AM) model of American English intonation: the existence of distinct phonological contrasts among nuclear tunes composed of a pitch accent (here H*, L+H*, L*+H), phrase accent (H-, L-) and boundary tone (H%, L%), which in combination yield an inventory of 12 tonally distinct nuclear tunes. Using an imitative speech production paradigm and AX discrimination task with L1 speakers of Mainstream American English (MAE) we test the extent to which each of 12 predicted tunes is distinct from the others in the production and perception of intonation. We tackle this question with a series of analytical methods. We use GAMM modeling of time-series F0 trajectories to test for differences among all of the twelve nuclear tunes, and compare these results to a method that does not rely on pre-defined tune categories, k-means clustering for time-series data, to discover emergent classes of tunes in a "bottom-up" fashion. We complement these time-series analyses with an analysis of the temporal tonal center of gravity (TCoG) over the F0 trajectories of nuclear tunes to assess tonal timing distinctions and their relation to top-down tune classes (defined by the AM model) and bottom-up classes (emergent from clustering). Production results are further compared to perceptual discrimination responses, which together point to a hierarchy of distinctions among nuclear tunes: a set of primary tune distinctions are emergent in clustering and always distinct in perception. Other tune distinctions, although evident in top-down analyses of (labeled) F0 trajectories, are lost in emergent clusters, limited in magnitude and scope, and often confused in perception. Results are discussed in terms of implications for a theory of intonational phonology.

1. Introduction

A recurrent question in intonation research concerns the relationship between abstract and categorical intonational distinctions, on the one hand, and the continuous nature and dynamics of prosodic characteristics (especially F0) in speech, on the other. One widely accepted approach to this issue is formulated in autosegmental metrical (AM) theory. In the broadest terms, the theory proposes that F0 patterns at the phrase level can be described in terms of sequences of abstract and categorically distinct H(igh) and L(ow) tones which map onto F0 targets, and which associate in a predictable fashion with the segmental string (see e.g. Jun, 2005, 2014 for descriptions of systems in various languages formulated in the AM framework). In the prevalent AM model of "Mainstream" American English (MAE) intonation (Pierrehumbert, 1980; Beckman & Pierrehumbert, 1986), tones associate with metrically strong syllables to mark prominence (in which case they are called pitch accents, indicated with an asterisk, e.g., H*). Tones also associate with the edges of prosodic domains, marking the edge of a smaller 'intermediate' phrase, in which case they are called phrase accents (indicated as H- or L-), or a larger 'intonational' phrase, in which case they are called boundary tones (indicated as H% or L%). We use the term 'edge tones' to refer to the sequence of phrase accent and boundary tone at the end of an intonational phrase. In this model, the phonological make-up of the F0 contour over a stretch of speech corresponding to an intonational phrase is accordingly a sequence of pitch accent(s), phrase accent(s) and boundary tone. The

* Corresponding author at: Dugald Stewart Building – University of Edinburgh, 3 Charles Street, Edinburgh, Midlothian EH8 9AD, United Kingdom.
E-mail address: jeremy.steffman@ed.ac.uk (J. Steffman).

https://doi.org/10.1016/j.wocn.2024.101310
0095-4470/

phonetic implementation of tunes is described in terms of (only) F0: H and L define relative F0 targets within a speaker's pitch range, with interpolation providing F0 values over regions that lack tonal specification. The sequence of a phrase-final pitch accent, phrase accent, and boundary tone is referred to as the "nuclear tune" due to the fact that the pitch accent marks obligatory phrase-level ("nuclear") prominence. Every intonational phrase thus contains a nuclear tune in the AM model, and it may also include optional pre-nuclear pitch accents and the possibility of additional phrase accents if made up of multiple intermediate phrases.

In what follows we refer to this model of MAE intonation as "the AM model". A core feature of the AM model is that phonological/categorical distinctions among nuclear tunes are the result of differences in (phonologically distinct) pitch accent, phrase accent, and boundary tones. All possible combinations of these phonological units generate a set of nuclear tunes and corresponding F0 trajectories in the language. The model thus defines a set of tunes that are distinct from one another in phonological units (tones), and which are associated with distinct F0 trajectories as produced by native speakers and which are heard as distinct by native listeners. In other words, the model is explicit in defining the contrast space of phonologically distinct forms, and the corresponding distinctions in the space of F0 trajectories. The present study focuses on three pitch accents, H*, L+H*, L*+H,[1] which have been subject to substantial debate in the literature. All three pitch accents contain a H(igh) tone and engender a pattern of rising F0. Combining these three pitch accents with phrase accents and boundary tones, the AM model generates a set of 12 phonologically distinct nuclear tunes (3 pitch accents $\times$ 2 phrase accents $\times$ 2 boundary tones), all of which are included in the tune inventory proposed for American English (Pierrehumbert, 1980; Beckman & Pierrehumbert, 1986).

The AM model has been widely adopted as a standard for the study of intonation in American English, though relatively little work has provided a rigorous test of its core assumptions, or empirical validation that the large number of predicted tune distinctions are recoverable from F0 signals. The literature on intonational meaning offers support for some of the distinctions proposed in the AM model, in showing associations between a particular tune and the pragmatic meaning of an utterance related to illocutionary force (speech act), or the speaker's epistemic state. However, such studies typically investigate a distinction between a limited number of tunes, often only two. Further complicating the picture, many studies of intonational meaning do not offer a clear description of the tunes under investigation, either in terms of a phonological (tonal) specification or the F0 characteristics, which limits their usefulness in testing the predictions from the AM model. Pierrehumbert & Hirschberg (1990) offer a more comprehensive account of intonational meaning for pitch accents, phrase accents and boundary tones, however there is to date no empirical study that validates each of the predicted tune-meaning distinctions from that work. An added challenge is the claim that pragmatic and discourse meaning may be conveyed by gradient phonetic variation, instead of or in addition to categorical distinctions in intonational form (e.g., Calhoun, 2012, Ladd, 2022). The intonational meaning literature thus supports the existence of contrasts among some tunes as conveying distinction in pragmatic meaning (e.g., Hirschberg, 2004; Prieto, 2015; Westera, Goodhue & Gussenhoven, 2020), though not for all of the tunes generated by the model and claimed to be part of the inventory.

The aforementioned challenge for the AM model concerns the evidence for discrete categories as a function of their pragmatic and discourse functions, but there also exists a related, perhaps more fundamental challenge to the model from the now well-documented variation in phonetic form (in F0 in particular), which calls into question the discreteness of proposed intonational distinctions. In a previous study, Cole, Steffman, Shattuck-Hufnagel, & Tilsen, 2023 explored this issue, using the same basic experimental methodology we use here (described below). In that study, the authors were concerned with the 8 nuclear tunes created from the combination of high (H*) and low (L*) pitch accents, with all boundary tone configurations (2 pitch accents $\times$ 2 phrase accents $\times$ 2 boundary tones). Using clustering and neural-net classification analyses over time-series measures of F0 in the tunes, Cole et al. (2023) found that three pairs of tunes were highly confusable: {H*L-H%, H*L-L%},{L*H-L%, L*L-H%},{H*H-H%, H*H-L%}. The two tunes in each pair were grouped together in the clustering analysis, resulting in only five emergent clusters from the predicted set of eight. The tunes in the same three pairs were also often confused in the neural net classification of the data. However, in other more targeted analyses of the same data, the authors found that there were measurable differences in the F0 trajectories of tunes that clustered together, though they were relatively small, and clearly insufficient to define separate clusters or to promote accuracy in the classification of the confusable tunes.

At a general level, the lack of robust distinctions in productions of certain proposed-to-be-distinct tunes highlights the issue of phonetic (F0) variability and intonational discreteness. However, one notable distinction evident in that study was between pitch accents H* and L*: tunes that differed in their pitch accent never clustered together, were successfully distinguished by the classifier, and were reliably perceived as distinct by listeners. Thus, while Cole et al. (2023) calls into question distinctions pertaining to edge tone configuration in nuclear tunes, their data offers clear support for a robust difference between high (H*) and low (L*) pitch accents. Our review of the literature (summarized in part below) suggests that the supposed difference between H* and L* pitch accents is indeed uncontroversial. Conversely, a clearly more contentious distinction is that between pitch accents that contain a high (H) target, and which vary in F0 alignment and scaling: H*, L+H* and L*+H (Beckman & Pierrehumbert, 1986; Silverman & Pierrehumbert, 1987). These distinctions, in particular, have been a longstanding topic of debate in the literature and their status as distinct intonational elements has been examined to some extent in previous studies.

[1] There are 5 pitch accents in the AM model's inventory for MAE: L*, H*, L+H*, L*+H, H+!H* (ignoring a downstepped !H*, which may occur only after another H). In these pitch accents the tone marked with a star is the one associated to the syllable with nuclear stress: for example L* +H and L+H* contain the same sequence of tones, but differ critically in how the L and H tones are associated with, or aligned to, the segmental string.

Cole et al. (2023) thus provides a jumping off point for the present study, in which we use a similar methodology to focus on H*, L+H* and L*+H.[2] Here we briefly review the way in which these pitch accents have been described as phonetically distinct. H* is a high target affiliated with the stressed syllable and is canonically described as containing a small rise to a peak within or after a stressed syllable, though as noted in the ToBI guidelines the actual timing of the peak can vary substantially (Beckman & Ayers Elam, 1997, p 15). The guidelines state that the critical distinction between H* and L+H* is the following: "...the essential difference [between L+H* and H*] is what happens before the high tone. The leading L tone in L+H* is meant to transcribe a rise from a fundamental frequency value low in the pitch range that cannot be attributed to a L* pitch accent on the preceding syllable or to a L- phrase accent or to a L% boundary tone at a preceding intermediate-phrase or intonation-phrase boundary. For H*, by contrast, there is at most a small rise from the middle of the speaker's voice range" (pp 15–16). Though not explicit in the guidelines, these accents may also be differentiated by F0 peak height, where for example, Burdin, Holliday & Reed (2022) found that L+H* shows a reliably higher F0 peak than H* in three varieties of American English. L+H* and L*+H are distinguished by the alignment of a low F0 region (often described as a low target) with respect to the stressed syllable. Whereas the L tone is at or preceding the onset of the stressed syllable in L+H*, a low F0 region occurs within the vocalic portion of the stressed syllable in L*+H, with a rise that starts late into or after the stressed syllable. In terms of F0 scaling, the L tone in L+H* has been shown to be higher than that in L*+H with a larger difference between L and H targets for L*+H (Arvaniti and Garding, 2007). To the extent that these three pitch accents are different phonological categories, we predict that each should evidence categorical behavior in production and perception, with F0 shapes that differ systematically in production and are reliably perceived as distinct by listeners. As noted by Dilley & Heffner (2013), despite the general acceptance of these categories, "little empirical evidence exists about how many accents truly underlie English intonation; for the most part, claimed distinctions have been based on descriptive evidence and theoretical arguments" (pp 7–8), a statement which, in our assessment, remains true almost a decade later.

There has been considerable debate in the literature relating to the claim from the AM model that H* and L+H* are two distinct accent categories (e.g., Calhoun, 2004, 2012; Ladd, 2008, Ladd & Schepman, 2003). Ladd and Schepman (2003) argue for a single category based on the existence of a lowered F0 which often precedes H* pitch accents, described in earlier work as a "sagging transition", and which behaves like a low target in that its alignment is conditioned by phonological factors and is relatively stable with respect to the following H tone. The alignment of the L target is also a perceptual cue to the presence of a word boundary for listeners. In their analysis, this low target constitutes a part of a (L + H) pitch accent, which also includes L+H*. In this analysis, there is a distinction in the span of the F0 rise, which varies along a continuum with canonical H* and L+H* as endpoints (cf. Calhoun, 2012, Dilley, 2005). Interestingly, these fundamental issues remain when discourse function and meaning are considered. It has been claimed that H* and L+H* accents are distinguishable in terms of their discourse function (e.g., Büring, 1997; Steedman, 2000). However, Watson, Tanenhaus & Gunlogson (2008), showed that while it is true that L+H* is interpreted as conveying contrastive focus, H* was also compatible with this interpretation. H* further was compatible with the presentation of a discourse-new referent, such that the discourse functions of the two pitch accents were not identical but were overlapping. In other words, interpretation in discourse context was not sufficient to distinguish the two pitch accents. This finding highlights that the mapping from pitch accent (if presumed to be a discrete category) to meaning/function is not one-to-one (see also Chodroff & Cole, 2019b; Im Cole & Baumann, 2023). Though not explored in Watson et al. (2008), these overlapping interpretations could also be seen as related to within-category phonetic parameter variation for a single accent category.

Though perhaps a lesser focus in the literature, the distinction between L+H* and L*+H has also been a topic of investigation. Arvaniti and Garding (2007) present evidence that these two pitch accents have largely different temporal alignment patterns, which is taken as evidence for two categories. However, one caveat in that study is that each pitch accent was elicited in the context of a different edge tone sequence: L+H* in the context of L-L%, and L*+H in the context of L-H% or H-L%. The distinctions in measured F0 parameters thus co-occur with variation in nuclear tune shape overall.[3] Although the differences in temporal alignment between the pitch accents are convincing, it is difficult to make generalizations about their distinctiveness *across* edge tone contexts, or when controlling for edge tones. In earlier work, Pierrehumbert & Steele (1989) carried out an imitative production task in which a peak alignment continuum between L+H* and L*+H was imitated in the context of an L-H% edge tone sequence. Imitations from three speakers showed a bimodal distribution of peak delay values, rather than the continuous variation along the steps of the peak delay continuum used as imitation stimuli. This is taken by the authors as evidence for a binary alignment distinction as the basis for the difference between L+H* and L*+H. Dilley & Heffner (2013) also examined the imitated productions of a low target continuum in a rising edge tone (H-H%) context. In the earliest alignment of the continuum, the low target preceded the stressed syllable, while in the latest alignment the rise occurred after the stressed syllable. Though described as an L* alignment continuum by the authors,

[2] In addition to being an empirical extension of Cole, Steffman, Shattuck-Hufnagel, & Tilsen, 2023 to different pitch accents and tunes, the present study also contains some different analytic methods, described in the methods section. One large addition is a focus on tonal timing and a more in-depth analysis of speech perception data. It should be noted that both this paper and Cole et al. (2023) make use of k-means clustering and GAMM modeling.

[3] Arvaniti and Garding (2007) further examined these differences in two dialects of American English. They found that speakers from Southern California exhibited overall later alignment of both high and low targets as compared to speakers from Minnesota, for both pitch accents, though relative timing differences between accents were preserved in both dialects. This shows the need to consider dialect-specific implementations of intonational elements, which in this case varied in the overall timing of tonal targets for the two pitch accents but could conceivably also influence relative differences between them. This constitutes an additional challenge for the understanding of claimed intonational categories to their phonetic manifestations in speech. The present study does not address dialectal variation, though we note that future work may build on these results to examine such questions.

the early alignment end point is similar to L+H*H-H%, with the continuum ranging from this to L*(+H)H-H%. The authors showed that listeners' perception of valley alignment in two continua of this sort was less accurate than that of peak alignment continua. They also found more graded variation in the production of valley alignment in an imitation task similar to that used in Pierrehumbert & Steele (1989). The data from Dilley & Heffner thus suggest a lack of a discrete distinction in F0 valley alignment in the context of an H-H% rising edge tone. The result from Dilley & Heffner, in comparison to the data from Pierrehumbert & Steele, shows that the evidence in favor of a categorical distinction is dependent on edge tone context. As with the H* versus L+H* distinction, the AM model for MAE proposes two categories, while other accounts of the distinction between L+H* and L*+H have invoked the notion of a phonetically gradient peak delay spanning the ranges between these two accents, with possible ideal or typical values (Gussenhoven, 1984). As stated by Pierrehumbert & Steele (p. 193), distinguishing between these accounts is not straightforward: "There is perhaps some conceptual difference between proposing two categories and proposing a gradient dimension with two preferred positions. However, it is unclear what experiments for deciding this could be realistically performed".[4] In another domain, speech technology research has also grappled with the issue of prosodic categories (e.g., Wightman & Campbell, 1995; Wightman & Ostendorf, 1994). Pitch accent categories are a recurrent theme here too. For example, findings of low inter-labeler reliability for ToBI labeling in text-to-speech applications led to collapsing among labels and the use of so-called "ToBI-Lite" annotation which only distinguished only two types of pitch accents (Syrdal & McGory, 2000; Wightman, Syrdal, Stemmer, Conkie & Beutnagel, 2000). Classifiers have also been developed to identify intonational features (e.g., Rosenberg, 2010; Schweitzer & Möbius, 2009). Setting aside the issue of labels within training corpora, confusions in pitch accent classification and detection are evident in American English (in e.g., Rosenberg, 2010). This generally lines up with classifier confusions for whole tunes seen in Cole et al. (2023).

In addition to the question of discreteness in tune form and function, a related issue merits consideration here. This might be called "probabilistic tone phonotactics", or the probabilities that certain tonal combinations will occur. The AM model proposes that all tone sequences are possible, but makes no claims about their relative frequency, which may be expected to impact the extent to which they are readily perceived and produced. In our reading of the literature remarkably little work has explored this question beyond Dainora (2001, 2006) who proposed a probabilistic grammar for intonational tunes based on the analysis of a ToBI-labelled corpus of radio news speech of two speakers. The grammar identified probabilistic dependencies between sequences of pitch accents, phrase accents, and boundary tones in nuclear tunes. Some, like H*L-L%, were very common, while others, like L*+H H-H% never occurred in the corpus. In the context of the literature reviewed above and the present study, this finding can be considered in two lights. First, and importantly, the probabilistic grammar takes the ToBI labels at face value and presumes that they represent discrete tones which combine to form tunes. Thus, unlike the questions of gradience raised above, this work presumes the existence of these tones as phonological units. Nevertheless, the key finding that tone sequences are highly variable in their frequency raises the possibility that the attested frequency of a tune may relate to its separability in production or perception. In other words, a particular pitch accent may be frequently realized in a given tune (H*L-L%) but not another (e.g., H*H-H% in Dainora, 2006), and this in turn may predict the extent to which it is distinguished from other pitch accents across contexts. This is an issue we return to in the discussion of the results presented here.

In sum, the literature offers mixed evidence for some of the predicted category distinctions of the AM model for American English. A critical gap is the generalizability of a proposed distinction across intonational edge tone (phrase accent + boundary tone) contexts (also highlighted by frequency discrepancies for certain pitch accent and edge tone combinations shown in Dainora, 2001, 2006). In each of the perception and imitation experiments reviewed above, the edge tone context was held constant, with potentially important differences suggested by the comparison between Pierrehumbert and Steele (1989) and Dilley and Heffner (2013). Moreover, covariation of pitch accent and boundary tone in Arvaniti and Garding (2007) makes apparent the need to examine pitch accent-based distinctions across edge tone contexts (and conversely, edge tone distinctions across pitch accent contexts). Building on these previous studies of categories and continua in intonational form, the present study is a test of the proposed phonological tune distinctions in American English, as an empirical test of the AM model. We test the predicted 12-way distinction between tunes constructed from one of three rising pitch accents H*, L+H* and L*+H with all possible edge tone combinations, with a large sample of speakers of the language. By testing the existence and nature of these distinctions we take a first step towards testing the range of nuclear tune shapes that are available as discrete categories in production and perception, and the scope of variation between and within them.

1.1. The present study

To elicit productions of American English intonation we carried out an imitation study, which follows previous work in using imitation as a method to gain insight into the phonological status of intonational features (e.g., Braun, Kochanski, Grabe, & Rosner, 2006; Chodroff & Cole, 2019a; Dilley & Heffner, 2013; Cole and Shattuck-Hufnagel, 2011). An alternative approach for eliciting intonational tunes would be to provide an explicit discourse context to elicit the production of an individual tune. However, the lack of a comprehensive understanding of the relationship between tune and discourse context for the full set of tunes we are interested in, as outlined above, makes this challenging. Moreover, previous empirical work investigating the meaning of rising intonation also focuses on only a small set of tunes, or a single aspect of pragmatic meaning (e.g., Burdin & Tyler, 2018; Jeong, 2018; Rudin, 2022). As such, we lack a clear basis for creating reliable discourse information for the 12 tunes of interest in this study.

[4] Some perceptual evaluations of categoricity have been carried out for edge tones, using identification and discrimination tasks for F0 continua (Remijsen & van Heuven, 1999; Schneider & Lintfert, 2003) though the crux of the debate has been focused around pitch accents in our reading of the literature.

We assume that tune imitation involves, as a first step, an encoding of its phonetic properties and associated phonological categories, which are subsequently implemented by the speaker in their imitative production of the tune. An imitation task thus allows us to focus on the phonological and phonetic form of the tunes in the AM model, which will provide a basis for future research to examine the mapping from tune to discourse context, and the pragmatic function of tunes. This should be kept in mind throughout as a caveat of the present study, which, for the reasons noted above, is a first step towards a broader goal of establishing the system of intonational contrasts in AE. The tunes in this study were presented without discourse context, as a test of the hypothesis that speakers and listeners should be able to access phonological tune categories in the absence of a meaningful discourse, in at least some of the phonological contexts in which they may occur. For instance, to support a categorical distinction between H* and L+H* pitch accents in nuclear position, we seek evidence that they are reliably produced and perceived as distinct in at least one context of a following phrase accent and boundary tone. Our decision to test tune distinctiveness in the absence of a discourse context is motivated on analogy with segmental phonological contrasts; while a contrastive relationship between features, gestures or phones may be marginal or even reduced in certain phonological contexts, contrastive status rests on there being at least one phonological context in which the distinction manifests in production and is recognized in perception.

Previous studies on the imitation of intonation show that not all aspects of a stimulus F0 contour are imitated; and that imitative productions are, in this sense, not a veridical rendition of the stimulus F0. For example, as reviewed above, it seems the alignment of F0 peaks is imitated consistently (with possible category-based discontinuities), but F0 velocity and (vertical) scaling along a continuum of low to high F0 values is not (Dilley & Heffner, 2013; Dilley, 2010; Tilsen, Burgess & Lantz, 2013). Imitation thus offers a window into the mental representations that are involved in the imitation process: the encoding of a tune in perception and the reproduction of that encoded information in production. Braun et al. (2006) showed that iterated imitation of randomly varying phrasal F0 patterns leads to the emergence of stable attractor patterns that correspond to predicted intonational distinctions, while also showing substantial within-attractor variability. In this sense imitation data also has the potential to capture variation within a category as well (cf. Calhoun, 2012, Ladd, 2022). An imitation task thus allows us to test distinctions among tunes absent an understanding of their discourse function and offers a test of the existence of discrete tune categories and possible variation within them. In the present imitation task we additionally sought to diminish the role of stimulus-specific episodic detail in several ways. In each trial (described in more detail below) participants heard two auditory models which presented the same intonational tune in two lexically different, though syntactically and metrically similar, sentences. Participants were then prompted to reproduce the tune on a new sentence using the same intonational melody they heard, such that the production is not, strictly speaking, an imitation of the stimulus. Each trial also presented stimuli from two different model speakers who had different pitch ranges, providing the participant with acoustically different realizations of the same tune. We complement this imitation study with a perceptual discrimination task testing discrimination for all pairwise combinations among the 12 nuclear tunes that comprise our stimuli. With these methodologies we examine the core questions outlined above. Our specific research questions can be summarized as follows:

1. How well does the AM model predict the observed distinctions in the shape of nuclear F0 trajectories produced by speakers as imitations of the 12 nuclear tunes formed using the rising pitch accents H*, L+H* and L*+H?

2. What acoustic parameters characterize the observed F0 distinctions in imitated productions, within and across the categorically distinct tunes proposed in the AM model?

Given the preceding discussion and these research questions, it is important to consider here precisely what categories are being sought, and how the present study aims to contribute to our understanding of intonational phonology. First, as stated above, the present study is not concerned with intonational meaning. It is fundamentally a study of intonational *form*, examining the intonational forms that speakers produce and perceive as distinct. This distinctiveness can be considered at two levels of description, the level of the tone (a phonological unit) and the tune (a more complex phonological structure composed of sequenced tones). Our analyses focus on the nuclear tune as a whole, given our interest in how intonational features (e.g., pitch accents) are distinguished in context (i.e., within a tune). To the extent that two tunes differing in a single intonational feature are distinct in production and/or perception, we have evidence supporting the categorical status of that intonational feature (as a phonological unit). For example, if L+H* and L*+H were distinct (in production and/or perception) in all edge tone contexts, this would qualify as a robust and cross-context distinction in intonational form. Conversely, if they are only marginally distinguished in few contexts, or are not well distinguished in any, this calls into question their status as unique intonational forms, i.e. as intonational categories. If they are robustly distinct in only some contexts, this could be viewed through the lens of contextual neutralization, a point which we return to in the discussion.[5]

Following this line of reasoning, one key part of our analysis is concerned with the intonational form distinctions that are *emergent* in the data, that is, the distinctions that appear from an analysis that is blind to the AM labels of the model tunes that speakers were imitating. We implement clustering analysis with this goal in mind, to partition speakers' productions into clusters on the basis of how distinct they are from one another (described in detail below). The clustering analysis reveals the distinctions that emerge from the data without recourse to information about the AM labels of the model, and in our view constitutes the most appropriate approach for assessing tune distinctions without assuming a set of *a priori* tune categories. In addition to the bottom-up, data-driven clustering analysis, we carry out additional analyses to compare the F0 trajectories

[5] In making reference to contextual neutralization, which is often associated with segmental phonology, we do not wish to imply that intonational tones, or tunes composed of strings of tones, should be assumed to behave like segments in general, or to be subject to the same constraints in production and perception. Some phenomena may be shared across domains however, for example, contextual variability and neutralization. Examining these for intonational tunes is one of the goals of the present study.

of imitated productions when identified with the AM tune label of the corresponding stimulus. These analyses of labeled data are AM-model-driven (or top-down, in contrast to the bottom-up clustering analysis). They allow us to assess the extent to which speakers imitate what they have heard, how their imitations deviate from the F0 trajectories of the model tunes presented as stimuli, and if imitations of different AM tunes can be distinguished from one another at all. We also relate these analyses of labeled data to the emergent clusters. However, unlike the clustering analysis, we do not consider the analyses of labeled data as the best method to answer the question of how many and which distinctions are *emergent* in the data. The analyses of labeled data assign *a priori* AM labels to imitated productions on the basis of distinctions present in the stimuli (by design), comparing F0 trajectories of imitations on the basis of those labels as a complement to the clustering analysis, which is neutral in regard to the AM labels. In summary then, the present study sets out to answer the question of which intonation forms are well separated in the production and perception of nuclear tunes, and the extent to which these comport with the 12 proposed nuclear tunes of interest. Implications of these results for intonational phonology, and intonational meaning are broached in the discussion section.

2. Methods

The model stimuli, the data reported in this paper, and the scripts for analysis and data visualization may be accessed from an open-access repository hosted on the OSF at https://osf.io/ehx7w/.

2.1. Stimuli

The speech materials used to create the stimuli were spoken by one male and one female speaker of American English. From these we synthesized twelve nuclear tune F0 trajectories using the PSOLA method and a custom Praat script (Moulines, & Charpentier, 1990; Boersma & Weenik, 2019). The two sentences were “He answered Jeremy” and “Her name is Marilyn”. The files were recorded in a sound-attenuated booth, using a Shure SM81 Condenser Handheld Microphone and Pop Filter, with a sampling rate of 44.1 kHz. Table 1 shows elicited nuclear word durations produced by each talker from aggregated productions of each word in the 12 nuclear tune conditions. The table also shows the duration of the selected base file for resynthesis. This selected base file had relatively low and flat F0 for both speakers, which was judged to sound the most natural with the range of resynthesized tunes.

In creating the F0 patterns for the 12 tunes our goals were two-fold. First, we wanted tunes to be an accurate reflection of the proposed distinctions from the AM model, for which targets and turning points correspond to the canonical description of each. To this end, the F0 patterns were created based on schema in the ToBI training materials (Beckman & Ayers, 1997; Veilleux, Shattuck-Hufnagel & Brugos, 2006), which in turn are based on schema in Pierrehumbert (1980). Our second goal was to ensure that each of the 12 tunes was maximally distinct from the others in F0 space. In other words, we wanted to ensure that the loss of a given distinction in imitative productions was not due to the stimuli not having sufficiently distinct F0 trajectories. All resynthesized tunes were judged by the authors (all ToBI-trained prosody researchers) to sound like plausible productions of the tunes, which were perceptually distinct. Notably, Cole et al. (2023) also elicited productions of H* tunes, though the H* tunes used in this study are slightly different than those in Cole et al., to maximize differences with L+H* (not a concern in Cole et al.; cf. Fig. 1 in this paper and Fig. 1 in Cole et al.). There is no overlap in the analyzed data between that study and this one, which recruited a fully new set of participants.

The resynthesized tunes were created based on 6 target heights distributed within each model speaker’s pitch range. The mapping of target height to F0 in Hz is shown in Table 2. Each model speaker had different target heights, with maxima and minima approximately matching natural variation in their pitch range, as determined from the model speakers’ natural renditions of each of the tunes. Intermediate targets were set to allow for fairly even spacing within the pitch range, with some values adjusted to improve the naturalness and the distinctiveness of the tunes, following perceptual evaluation by the three authors of this paper. The 12 tunes as represented schematically by F0 target heights are shown in Fig. 1. In describing the stimuli and tunes in the remainder of the paper we use a tune label without diacritics (+ - %) such that for example L*+H H-H% is displayed as L*HHH, L+H* H-H% is displayed as LH*HH, and so on. The timing of targets was based on temporal landmarks corresponding to syllable boundaries that were identified auditorily and by visually inspecting spectral transitions in the sound files. One additional temporal landmark in the nuclear region was used for L*HLH (L*+H L- H%), which necessitated an additional turning point. This additional landmark was set to one third of the duration through the final syllable (see Fig. 1), which based on auditory assessment of the three authors, sounded the most natural compared to earlier or later timing in the final syllable. Note that the ToBI training materials only provide alignment information with respect to the stressed syllable, and do not provide additional alignment landmarks. Alignment of the second turning point in Fig. 1 (rightmost dashed vertical line in each panel), was established as the transition between the

Table 1
Durational information (in ms) for nuclear word in the files which were recorded in the process of stimulus creation. The duration of the file which was selected to be the based file for the creation of all resynthesized tunes is shown at right.

Speaker	Nuclear word	mean(sd)	min	max	selected base file
female	Jeremy	547(41)	421	609	560
female	Marilyn	587(40)	529	639	527
male	Jeremy	649(65)	550	770	633
male	Marilyn	654(71)	561	770	712

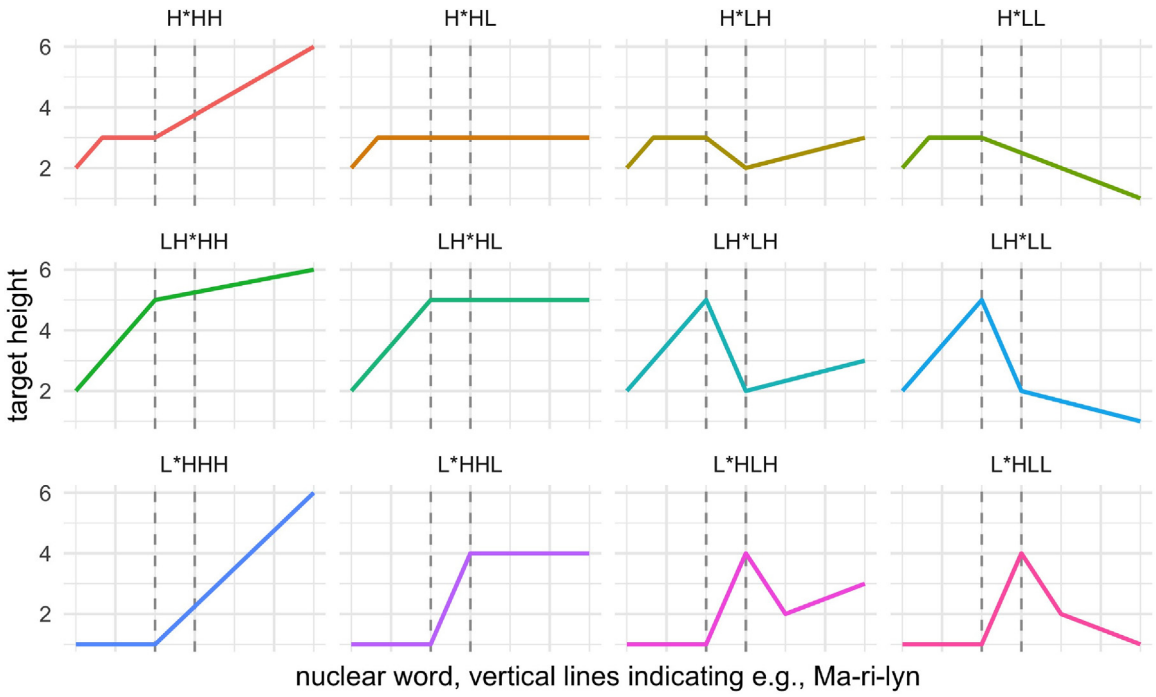

Fig. 1. Schematic representations of the model stimuli in the nuclear (final) word, with vertical lines indicating alignment with segmental material in terms of syllable boundaries.

Table 2
Hz and ERB values for the targets used in the model stimuli.

	Male model speaker		Female model speaker	
	Hz	ERB	Hz	ERB
Target level 1	80	2.79	100	3.37
Target level 2	105	3.51	160	4.93
Target level 3	130	4.18	200	5.84
Target level 4	225	6.36	300	7.79
Target level 5	240	6.67	350	8.62
Target level 6	265	7.15	380	9.09

second and third syllable in the nuclear word, as this location was judged to sound natural and to be systematically identifiable for the resynthesis.

In addition to the variation in the nuclear region shown in Fig. 1, we varied F0 in the preceding words, which we refer to as the "preamble". For a given pitch accent, the preamble was always the same across the two stimulus sentences, but the preamble varied by pitch accent. For the H* accent the preamble stayed at the target 2 level throughout until the nuclear word. For the L+H* accent the preamble started 20 Hz above the target 3 value and fell linearly to the target 2 value at the start of the nuclear word (note that this 20 Hz was the same value for both the female and male model speaker, and thus may be perceived slightly differently based on the different pitch ranges in the stimuli). The preamble for the L*+H accent started at this same point and fell to the target 1 value at the beginning of the nuclear word for that pitch accent. These additional manipulations of the preamble were found to be required for the onset of each nuclear region to start at a level that sounded natural, and to make the nuclear pitch accent more natural in relation to the preamble. The trajectories for the 12 tunes, including the preamble, are given in the appendix in Fig. A1.

2.2. Speech production experiment

2.2.1. Participants

We recruited a total of 70 speakers for the experiment, all of whom were self-reported monolingual speakers of American English, with no hearing deficits. Participants were recruited from two platforms. 35 participants were recruited from the subject pool at Northwestern University (17 women, 17 men, 1 non-binary, mean age = 19.7), and received course credit for their participation. The additional 35 speakers were recruited from Prolific (19 women, 14 men, 2 non-binary, mean age = 23.7), and were paid for their time. A preliminary analysis did not find any systematic differences between these two groups' tune productions. We carried out this analysis by fitting a GAMM model to trajectories for each tune, for each group (same structure as GAMM described below, but split by group as well). We then compared the production of a given tune across groups by examining difference smooths. This effectively compares if/where along a trajectory in normalized time, the groups differ from one another in their production of a particular tune. Examination of difference smooths revealed no significant differences in tune production between groups for 11 of the 12 tunes (no region of significant difference in the dif-

ference smooths). For one tune, L*HHH, there was a small region where F0 in the rising F0 movement was slightly lower for the group recruited from the Northwestern University subject pool. However, the difference was very small in magnitude and only persisted for approximately 20% of the tune (from the interval of approximately 60%–80% in normalized time), leading us to conclude it did not constitute a substantial difference across groups. This, in combination, with *no* difference detected for the other 11 tunes, led us to combine the groups in the remainder of the analysis. We further performed a qualitative screening of participants' productions to exclude any participant who did not produce substantial F0 variation during the experiment, i.e. a consistent monotone production across trials, which would reflect a misunderstanding of, or inattentiveness to, the task. This screening was accomplished simply by plotting each speaker's F0 trajectories for all trials, and inspecting them. We found that all speakers produced substantial variation across trials, with a particularly clear visual separation between monotonically rising and rising-falling F0 shapes, leading us to retain all participants for data analysis. A figure included online in the supplementary materials shows these by-speaker plots (averaged by tune for visual clarity).

2.2.2. Procedure

Participants completed the experiment remotely and were instructed to be seated in a quiet room and to wear a pair of headphones during the experiment, with recording done using their own computer hardware. The experiment started with a headphone check during which participants were presented with audio (piano music). Participants confirmed that they could hear the audio, and set their headphone volume to a comfortable level. There was an additional microphone check in which participants were prompted to speak and were shown a dynamic registration of the volume of their speech, to confirm that it was being recorded and was not too loud or quiet.

Participants were told that they would hear "computer generated speech" and that they should listen to the model utterances for a given trial and then produce a new sentence with the same melody "said the way you think it should sound if it were spoken by a human English speaker". This was intended to allow for participants to modify or enhance their production of the tune, and not focus on producing a phonetically precise imitation. A trial consisted of the auditory and orthographic presentation of two model sentences, each with the same nuclear tune, separated by one second. We refer to these as the *exposure* tunes for a given trial. The model sentences varied both in model speaker gender, and model sentence order (i.e., a participant never heard the same speaker gender or sentence twice in the same trial). With this design there are thus four possible configurations for the model stimuli (2 speaker orders $\times$ 2 sentence orders), which were paired with each of the 12 tunes for a total of 48 unique model stimulus presentations (i.e., 48 unique exposure tune conditions). These were matched with the three target sentences (produced by the participant) "He modeled harmony", "She remained with Madelyn" and "They honored Melanie" for a total of 144 trials. The target sentence was presented orthographically, with a reminder text that read "I would say it this way". The experiment took approximately 25 minutes to complete.

2.3. Speech perception experiment

To examine the perceptual discriminability of these tunes, we carried out an AX discrimination experiment in which listeners were presented with pairs of tunes from the pitch-resynthesized model stimuli and responded with a "same"/"different" decision. We opted for an AX paradigm in lieu of an AXB or ABX design largely because we assumed that three full sentences in a trial (six syllables per sentence) would induce high memory demands and be a challenge for listeners to store and compare each stimulus. Because our stimuli were sentences (not single syllables), the simplicity of the AX task is appealing in this regard. Use of the AX paradigm, however, does have the drawback that "different" judgments may only be given in cases of high certainty, with the relative differences across trials establishing a benchmark for the range of stimulus variability (see e.g., Gerrits & Schouten, 2004 for discussion). In this sense, a "same" response may reflect uncertainty about a subtle difference, though we assume that, when aggregated across participants, the rate of "same" responses should index the relative salience of differences among pairs of tunes.

In each trial, two model stimuli were played in succession, separated by an inter-stimulus interval of 500 ms. They were matched in terms of the model speaker gender and the model sentence, such that in a given trial only the tune (F0) varied. Tunes were paired with one another in all possible order-sensitive combinations (12 $\times$ 12) for a total of 144 pairs, including the pairing of the tune with itself (a "same" trial). To increase the proportion of same trials we doubled them, for a total of 24 same trials, and 156 total pairings (132 different trials, 24 same trials). These 156 parings constituted the 156 trials in the experiment.

Recall that for a given tune there are four files: two model speakers and two sentences. Including all combinations of these four files with the tune pair manipulation would have greatly increased the number of trials in the experiment. Accordingly, four different lists that contained the 156 pairings were created. Each list counterbalanced model speaker gender and sentence such that across the four lists all possible combinations were attested, and within a list there was an equal number of stimuli with each model speaker gender and model sentence. Over the course of the experiment, participants thus heard both model speakers and both sentences, though not every possible unique combination of these and tune pair. We recruited a total of 60 participants (44 women, 14 men, 2 non-binary, mean age = 28.7) and assigned 15 participants to each of the four lists, such that, across participants, every possible combination of model speaker, sentence and tune pair was attested. None of the participants from the speech perception experiment took part in the speech production experiment.

All participants were self-reported monolingual speakers of American English recruited from Prolific and were paid for their

time. The experiment took approximately 15–20 minutes to complete. There was a headphone check at the start of the experiment, as in the imitative speech production experiment.

2.4. Analyses

To address the questions laid out above we carry out complementary analyses of speech perception and production data. We will begin the results section with the inspection of the speech perception data and its relation to phonetic differences between pairs of model tunes. This will allow us to characterize the extent to which measurable F0 differences in the stimuli used for both perception and production analyses were perceived. We then turn to the production data, using GAMM modeling. Building on these results we carry out the "bottom up" clustering analysis, which is blind to the tune label associated with the stimulus (based on the AM model) for a given imitated production. This defines emergent distinctions between the imitations based on differences between them in F0 space. We then assess distinctions among imitated tunes in terms of their Tonal Center of Gravity (Barnes, Veilleux, Brugos & Shattuck-Hufnagel, 2012; Barnes, Brugos, Veilleux & Shattuck-Hufnagel, 2020).

As noted above, whereas the "bottom-up" clustering approach identifies an optimal partition of the data that is blind to AM tune labels, the "top-down" modeling approach with GAMMs and Tonal Center of Gravity allows us to relate these emergent cluster distinctions to the AM model. At a basic level, the two "top-down" approaches also allow us to test if a particular pair of AM tunes can be distinguished at all, e.g. it may be the case that H*LL and LH*LL cluster together, but they have some detectable difference in either Tonal Center of Gravity or in the time-series trajectories submitted to the GAMM analysis. Relating these two approaches to data analysis thus allows us to assess claims of categoricity from both perspectives and to test for the extent to which AM-label-based distinctions are preserved both within and across emergent clusters.

Finally, we will return to the speech perception data, examining how listeners' perceptual discrimination of tunes can be related to the speech production results. In the remainder of this section we give the details and motivation for each of the analyses.

The analyses presented here examine phonological tunes in relation to a single acoustic parameter, F0. In focusing exclusively on F0, we follow a large body of prior research on intonational phonology that takes F0 as a primary (or only) object of analysis, though in doing so we do not intend to suggest that other cues are unimportant for conveying distinctions among tunes. Two obvious candidates to consider are duration and voice quality. Our decision to limit our focus to F0 is based in part on the fact that in the phonological model we are testing, the AM model, tunes are characterized solely in terms of F0 targets and trajectories. An additional consideration is that further empirical work on how other acoustic parameters vary between and within tunes is needed before those parameters can be included as co-variates of F0 in experiments such as we present here.

2.4.1. Measurement and data processing

Audio files were force-aligned using the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner & Sonderegger, 2017). Each file was subsequently checked by trained auditors, who corrected alignment in the nuclear word and the alignment of phone boundaries from the phone tier (used to indicate syllable boundaries in the nuclear word) when needed. In this process files that contained disfluencies, misproductions of the target sentences or clipped or otherwise corrupted audio were excluded (3.7% of files). We extracted F0 using the STRAIGHT algorithm as implemented in VoiceSauce (Kawahara, Cheveigne, Banno, Takahashi & Irino, 2005; Shue, Keating, Vicenik & Yu, 2011), measuring F0 at every 10 ms across the file. In our analyses that deal in normalized time, these F0 measures were converted to 30 time-normalized samples across the nuclear word.

We then excluded files that contained unreliable F0 estimates. Because our measures are coming from a phrase-final word, we expected that non-modal phonation may be present which renders F0 estimation difficult (see e.g., Penney, Cox & Szakay, 2020). To identify files that contain F0 estimation errors, we used an algorithm that computes sample-to-sample differences in F0 (Steffman & Cole, 2022) and flags files with differences that exceed a threshold based on physiological thresholds from speech production research quantifying the maximum rate of change in F0 (Sundberg, 1973). The files that were flagged by the algorithm were subsequently inspected by trained auditors for possible re-inclusion in the data set. Files were re-included only if the F0 estimates did not show sudden discontinuities (i.e., upward or downward jumps) and comported with the perceived pitch in that region. In total 8.8% of the files were excluded using these criteria. We retained 8,913 files for analysis.

Speaker means and grand means for each tune are shown in normalized time in Fig. 2, plotting ERB scaled within speaker to normalize for different F0 heights and ranges. This can be compared to Fig. 1 to give a rough sense of how the model tunes were reproduced by our participants, and it can be remarked that the imitated productions overall correspond to the distinctions present in the model stimuli, though evidently with large variation across speakers.

2.4.2. Analysis of speech perception data

The modeling of speech perception data focused on "different trials", that is, trials in which the pitch-resynthesized tunes presented to the participant differed acoustically and in their phonological tune label ("same" trial results were essentially at ceiling in terms of accuracy). We modeled the perception results using Bayesian mixed effects logistic regression models (Bürkner, 2018). Several different models are presented and explained in the results section, however they all have the following in common. The binomial "same"/"different" response was predicted as a function of acoustic distance between the two tunes presented on a given trial, where distance was computed using the standard formula for root-mean square difference (RMSD). This metric has been shown to correspond well with perceptual ratings of intonational distinctiveness (e.g., Hermes, 1998), and offers a straightforward way of capturing overall differences between two F0 contours. The formula is given in (1) below:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{N}} \quad (1)$$

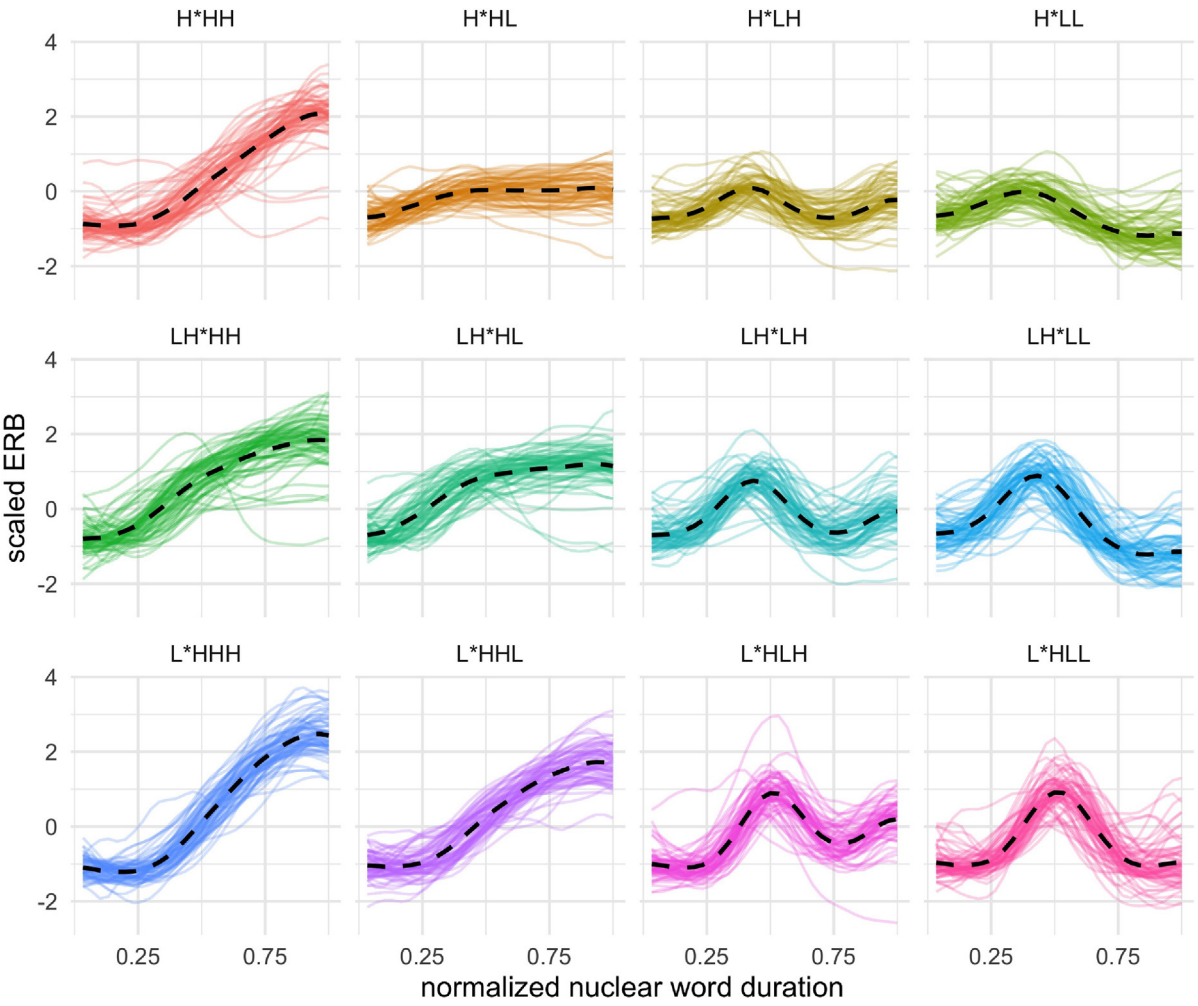

Fig. 2. Means for each tune (dashed lines) and each speaker (solid colored lines).

In (1), x_i is the ERB of one model tune at time n (for 30 time-normalized samples per nuclear word), and y_i is ERB of the other model tune at the corresponding normalized time.[6] Greater RMSD values indicate a greater distance between two model tunes in F0 space. For the purpose of statistical modeling only, the RMSD values are subsequently centered such that a value of zero is the mean RMSD for all tune pairs, which facilitated setting priors in the model. Importantly, we plot non-centered RMSD in the figures in this paper. Note that because each tune pair was presented in four different files (crossing the two model speaker genders with the two sentences) a given tune pair has four different RMSD values. Additional variables based on clustering results are included in some models of the perception data, which will be described in subsequent sections. For all perception models, random effects consisted of a random intercept for participant, and a random intercept for "base item", that is, the audio file from which all tunes were synthesized (one of four files, which crossed model speaker gender and sentence). To allow for the possibility that listeners may be influenced differently by variation in RMSD, a by-participant random slope for centered RMSD was additionally included in the model. In models with other fixed effects (described in subsequent sections), these effects were additionally included in the model as by-participant random slopes. Priors for the perception models were set to be normally distributed with a mean of 0 and standard deviation of 1.5 (in log-odds) for both the intercept and fixed effects. These can be considered "weakly informative" priors in that they encode no prior expectation of a fixed effect (with high uncertainty); the prior for the intercept is essentially flat in probability space (McElreath, 2020), allowing a wide range of possible intercepts.

[6] These were represented as speaker-centered ERB values (0 = speaker mean ERB), however, both centered and non-centered values render the exact same RMSD, as the differences between tunes does not change as a function of centering.

2.4.3. GAMM modeling

Our first analysis of the speech production (imitation) data makes use of a Generalized Additive Mixed Model (GAMM). The principal goal of this analysis will be to arrive at a characterization of the imitations according to the exposure tune label, and to see if there are measurable differences among all AM tunes. Though it is also possible to compute RMSD for pairwise combinations of tunes from the speech production data (like we did for the model stimuli), RMSD as a holistic measure would fail to capture *where in the F0 trajectory* differences exist and would be more difficult to compare across

tunes when dealing with many possible tune pairings. This led us to opt for using a GAMM, modeling F0 as a time-series over the nuclear word, to provide view of the imitations as labeled by exposure tune.

The GAMM model was fit to predict speaker-centered ERB over normalized time (in 30 samples), using the R packages *mgcv* and *itsadug* (van Rij, Wieling , Baayen & van Rijn, 2020; Wood, 2017). The GAMM was an AR1 error model, fit with parametric terms for tune category (coded with H*HH as the reference level[7]), and smooth terms for tune category over time. Random effects were specified as reference/difference smooths for speaker and tune and a random smooth for word, following Sóskuthy (2021). The default number of basis functions for each smooth term was found to be adequate, as determined by the *gam.check()* function. As the coefficients for the parametric or smooth terms are not particularly informative for the questions we ask here (Sóskuthy, 2021), we rely on visual inspection of GAMM fits and 95% confidence intervals to evaluate the significance of differences between tunes and to give a qualitative overview of the distinctions present in the data. The full model can be found on the open access repository.

2.4.4. Clustering

We assess emergent distinctions in imitative tune productions using k-means clustering for longitudinal data using the R package *kml* (Genolini, Alacoque, Sentenac, & Arnaud, 2015). The algorithm identifies distinctions among imitated F0 trajectories that are sufficiently large and robust to define distinct shape-based clusters. We inspect the output of the optimal clustering solution to determine how the emergent clusters relate to the twelve AM-defined model tunes, examining the proportion of imitations of a particular tune that are assigned to the same cluster. We also qualitatively describe the shape of the average trajectory for each emergent cluster. The clustering algorithm operates by comparing clustering solutions that are seeded using different numbers of clusters. Each clustering solution represents the optimal grouping of observations (i.e., F0 trajectories, each a vector of 30 F0 values) into *k* clusters.

One feature of k-means clustering that is different from other methods such as hierarchical clustering,[8] is the requirement that a specific *k* (number of clusters) be set at the start of the analysis. Choosing the best number of clusters for a particular data set thus becomes an important consideration in the analysis, and for this reason, clustering solution *optimization* constitutes a vital part of the k-means analysis. We accordingly ran the analysis with a range of *k*s, and then selected the optimal solution (using the functionalities in the *kml* package). We tested models using 2–12 *k*s (clusters), two representing the logical minimum number of clusters in the data set and 12 representing the total number of tunes in the input, which in theory could each define a separate cluster. The optimal solution in *k*-means clustering is defined as the partition of the data that minimizes within-cluster variance while maximizing between-cluster variance. Among these solutions, the optimal number of clusters was determined using the Caliński-Harabasz criterion (Caliński & Harabasz, 1974), to identify the value of *k* that results in the highest ratio of between- to within-cluster variance.[9] The clustering solution we report in the paper is the one that best separates the data, though it is only one of the eleven possible solutions (several additional solutions are mentioned in the results section, and contained as figures on the open access repository). We viewed this feature of k-means clustering, and the optimization component of the process, as a way to examine the category-like structure in the data, reasoning that emergent clusters in the data set are good candidates for categorical intonational (form) distinctions, and the number of clusters in the optimal solution represents the best category candidates based on cluster separation.[10]

The clustering analysis was carried out on speaker means for each tune (each speaker contributing 12 means), such that there were a total of 840 trajectories submitted to the algorithm. F0 was represented as ERB, and scaled within speaker, to remove differences in F0 height and range across speakers.

2.4.5. Modeling tonal timing using Tonal Center of Gravity

The GAMM modeling analysis presents a characterization of each imitated tune in terms of its dynamic F0 trajectory, allowing us to compare overall tune shapes, while the clustering analysis serves a similar purpose though without a prespecified category (i.e., tune label) for each tune.

Given that both the clustering analysis and GAMM modeling operate over time-normalized trajectories, a natural complement to these analyses is one that captures temporal distinctions among tunes in real time. As discussed above, a common approach to investigating temporal distinctions in intonation is in the analysis of peak and valley alignment (e.g., Dilley & Heffner, 2013; Ladd, Mennen & Schepman, 2000, Pierrehumbert & Steele,1989). For the present study, the alignment of F0 peaks and valleys is predicted to distinguish nuclear tunes, with, for example, the alignment of F0 peaks for an L*+H accent later than those for and L+H* accent. However, recent work in the intonation literature has called into question the primacy of turning points (peaks and valleys) in an intonation contour as the key parameter distinguishing pitch accent categories. The alternative *Tonal Center of Gravity* model (Barnes et al., 2012,2021) proposes another approach for investigating tonal timing. As described in Barnes et al. (2012), the Tonal Center of Gravity (TCoG) within some interval localizes an F0 event in time and in F0 space. There are two parameters that measure this localization. One is TCoG in the frequency domain, “frequency TCoG”, which can be

[7] Re-leveling the reference level will change the estimates for parametric terms in the model, but does not impact the visualization of the smooth terms which are presented and discussed in the results.

[8] Kaland (2021) provides an introduction to time-series F0 hierarchical clustering using the agglomerative method with complete linkage and accompanies an excellent user-friendly tool for the method.

[9] The Caliński-Harabasz criterion is defined by the ratio of between-cluster variance (distance between cluster centroids) to within-cluster variance (distance between an observation and the centroid of the cluster it is assigned to), also considering the number of observations and number of clusters. The optimum clustering solution is identified by a peak in the CH criterion calculated over an increasing number of clusters and corresponds to a solution with clusters that are both maximally dense and well-separated.

[10] In comparison, hierarchical clustering does not entail specification of *k* or optimization among clustering solutions that vary in the number of clusters. Instead, there is a single clustering solution, represented by a dendrogram, which may be “cut” in different locations to yield various numbers of clusters. Optimization among these different cuts of the solution can also be caried out in various ways. Given that k-means clustering explores and evaluates different solutions (different values of *k*), we thought it a more appropriate way to address our questions here. The open access repository contains visualization of hierarchical clustering analysis of the data, carried out using the tool described in Kaland, 2021; this is referenced below as well in light of the clustering results.

computed as the average F0 of the interval under consideration, or in a more complex fashion, can also incorporate weights that emphasize particular regions in the signal. The second parameter, which we focus on, is "temporal TCoG": the temporal location of TCoG within a particular interval. Given, e.g., an H* pitch accent, temporal TCoG measures when (in time) the F0 "bulk" or "mass" of the pitch accent occurs and can be computed with the equation given in the appendix. What the TCoG computation captures is a *holistic* characterization of where F0 is highest, along the course of an F0 movement. The TCoG measure is influenced by the shape (domed vs. scooped) of the F0 trajectory. Fig. A1 in the appendix provides an illustrative example of this effect, similar to examples given in Barnes et al. (2021). These sorts of shape-based distinctions have been attested in speech production and shown to be important in speech perception (Barnes et al., 2021; 't Hart, 1991; d'Imperio, 2000; Knight, 2008; Niebuhr, d'Imperio, Fivela, & Cangemi, 2011). Notably, the shape-based distinctions among rising tunes that are captured by the TCoG measure are not predicted by the AM model, which implements F0 interpolation between successive tone targets. Most relevant to the point made here, Barnes et al. (2021) show that shape modulations impact how listeners perceive the timing of an F0 peak (even when actual peak timing is fixed). Such results support the idea that the timing of tonal events is computed by listeners as something akin to TCoG, in lieu of actual peak timing. We take this result and related ones in the intonation literature to favor temporal TCoG as a measure of tonal timing in tunes, and we therefore apply this measure in our the search for predicted distinctions among them.

The TCoG analysis presented in this paper is aimed at exploring how TCoG relates to both the emergent distinctions in the clustering analysis and the AM tune labels. First, in similar fashion to the GAMM analysis, we can ask how TCoG differentiates imitations when labeled by AM model tune, i.e. are there detectable differences among each of tunes? Second, we ask: is tonal timing, as measured with TCoG, a parameter that captures the emergent clustering distinctions? If yes, we can infer that emergent partitions of the data are based (in part) on tonal timing. Finally, in relating both the clustering and AM-based distinctions, we will consider how TCoG predicts variation in a particular AM model tune when it is split across multiple clusters. The logic of these comparisons will be described in more depth in the results section below.

In implementing a TCoG analysis, a practical question is what interval should be selected for the purpose of computing TCoG, though it appears that TCoG measures are fairly robust to this choice (Barnes et al., 2012). We believe analyzing the nuclear tune as a unit offers a natural answer to this question, as it defines a linguistically meaningful interval that contains a pitch accent followed by a boundary marking event. In this case, a pitch accent with a later peak and a rising-falling F0 shape, e.g., L*+H L-L%, will have later TCoG within the nuclear tune interval than L+H* L-L%. Moreover, any shape-based variations in the tunes will be captured in this metric making it potentially superior to an analysis of F0 peak timing. Finding a systematic effect on TCoG for these tunes with a rising-falling F0 pattern would replicate findings in previous studies, where for example Barnes et al. (2012) show that L*+H entails a later TCoG than L +H*. Additionally, we believe that the usefulness of the TCoG model extends beyond the analysis of rising-falling F0 shapes. As will be shown, monotonically rising nuclear tunes also differ systematically in shape, some containing domed rises, others containing scooped rises. In the present study we utilize TCoG as a tool to quantify these shape-based distinctions in rising F0. This latter application of TCoG is, to our knowledge, a novel extension of its use as a parameter for quantifying differences in tonal timing.

TCoG was computed within the nuclear word, with F0 measured in ERB. We "anchored" the TCoG measurement to a landmark in the segmental string, namely the boundary between the first (pitch-accented) syllable, and the following syllable. This was accomplished by subtracting the timepoint at the end of the first syllable from the temporal TCoG value (in ms). This landmark struck us as a practically useful one in the sense that negative values of TCoG alignment will mean that the temporal TCoG is within the first syllable, and positive values will mean it is after the first syllable.[11] Secondly, this anchor point strikes us an important one considering the tunes under analysis: in the model tunes, the placement of the peak, or the rise varies systematically around this syllable. Statistical analysis of the TCoG data was carried out using Bayesian mixed-effects regression, fit to predict the dependent variable as a function of pitch accent and edge tone, and the interaction of these effects. The random effects in this model included intercepts for speaker, with by-speaker slopes for pitch accent, edge tone, and their interaction. Additional random intercepts were included for the word over which each tune was produced (harmony, Madelyn, Melanie), allowing for the possibility that the segmental material in each word influenced TCoG. We carry out an additional TCoG analysis informed by the clustering results, which is described in more detail in Section 3.4.

In reporting results from Bayesian models,[12] we provide the posterior median estimate and 95% credible intervals (CrI). When CrI exclude the value of zero, we take them to provide compelling evidence for an effect, i.e. a clearly non-zero effect size (see e.g., Vasishth, Nicenboim, Beckman, Li, & Kong, 2018). For the purposes of effect interpretation, we focus on effects that meet this criterion. We additionally report *pd*, computed with the *bayestestR* package (Makowski,Ben-Shachar,& Lüdecke, 2019). This metric gives the percentage of a posterior which shows a given directionality, ranging between 50% (a distribution centered exactly at zero: no effect) and 100% (a distribution which excludes zero entirely: a clear effect). The models and model summaries are included in the open access repository.

[11] Segmentation of the first/second syllable boundary based on acoustic landmarks was straightforward in the target words "har-mony" and "Ma-delyn" because the onset on [m] and [d]/[ɾ] was consistently clear. In the case of "Melanie", the F2 minimum was selected as the most readily identifiable acoustic landmark for [l], which means that the "syllable boundary" in this word is essentially the apex of F2 movement. This makes TCoG measures across target words slightly different in nature, though consistent within a word.

[12] In all of the Bayesian modeling we report here, the model was fit using a no U-Turn sampler with an *adapt_delta* parameter setting of 0.99. The models drew 4,000 samples in each of four chains, with a burn in period of 1,000 samples to ensure adequate independence of from the starting values of the Markov chains, effectively retaining 75% of the samples for inference.

3. Results

3.1. Perception data: Effect of RMSD

In this section we examine one analysis of the speech perception results, which will be referenced in the subsequent speech production analyses. Fig. 3 shows the perceptual discriminability for all tune pairs, where the y axis plots the proportion of a correct "different" response for different pairs in the perception experiment, with the specific tune pair labelled on each point.

One consideration in our modeling of perceptual discrimination responses was whether to include RMSD in the preamble (the two unaccented words preceding the nuclear word) as a predictor in the model. We compared two models, one with just nuclear RMSD and one with both nuclear and prenuclear (preamble) RMSD, plus the interaction of these two fixed effects. The model including preamble RMSD found that higher preamble RMSD in fact predicted *poorer* discrimination performance ($\widehat{\beta}$ = -0.28, 95% CrI = [-0.53,-0.04], pd = 99). This is likely related to the fact that, within a pitch accent, the preamble is identical (0 preamble RMSD), and yet tunes with the same pitch accent may vary largely as a function of edge tones.[13] Henceforth we model RMSD in the nuclear region only (the model with preamble RMSD can be found on the open access repository). The relationship between nuclear RMSD and preamble RMSD is shown in Fig. A3 in the Appendix.

Recall that the RMSD measure we use in modeling the perceptual discrimination results also crucially captures variation due to the model sentence (among two) and model speaker gender, as each tune pair RMSD measure was computed four times: once for each combination of model speaker and sentence in the stimulus set. We show how the tunes vary in both of these parameters in panels B-D in Fig. A3 in the Appendix, where it can be noted that there is a minimal influence of either sentence or model speaker on tune discrimination. In the remainder of the paper we thus focus on tune pair effects, though it should be kept in mind that four separate RMSD measures for each tune pair are input to the statistical model of the perception data.

Returning to Fig. 3, the key observation is the general relationship between pairwise RMSD and perceptual discrimination performance. This was confirmed in modeling, whereby there was a credible effect of (nuclear) RMSD on the log-odds of a correct "different" response: listeners are more accurate discriminating a pair of tunes on a given trial when the tunes have larger RMSD ($\widehat{\beta}$ = 0.72, 95% CrI = [0.63,0.81], pd = 100; this notably agrees with the estimated effect of nuclear RMSD in the model that also includes preamble RMSD, where $\widehat{\beta}$ = 0.76). This result shows that the phonetic distance between tunes, as computed with RMSD, goes far in predicting listeners' discrimination performance. However, we also observe a wide range of discrimination accuracy among tune pairs with roughly the same RMSD, particular around RMSD values of 2. We conclude that RMSD leaves substantial variability unaccounted for, suggesting there is additional information that is relevant to listeners' discrimination of tunes. In other words, if discrimination of tune pairs was strictly a function of the distance between tunes in F0 space, we would expect less variability in responses within a similar RMSD region. With this in mind, we turn to the speech production results.

3.2. GAMM fits: A first examination of tune imitation

The main goal of GAMM modeling was to assess the overall extent to which speakers produced different tunes, as labeled by the model (exposure) tune. Fig. 4 shows the GAMM fits for imitations of the 12 model tunes grouped by edge tones, with three trajectories in each panel showing the three pitch accents for that edge tone condition. Given that many pairwise comparisons are possible (66 in total) we do not step through each of these individually. Fig. A4 in the Appendix shows difference smooths for pairwise comparisons of interest.

Fig. 4A shows three trajectories for tunes that end in the HH (H-H%) edge tones, each of which shows the expected rising shape. A pitch accent distinction is also present among the three trajectories in this panel, with LH* showing an earlier rise than L*H. We also note a difference in the curvature of each rise, whereby LH*HH shows a more domed rise, and L*HHH shows a more scooped rise. H*HH shows an unexpected pattern in that its overall shape is similar to the trajectory with the L*H pitch accent: a later rise with a more scooped shape.[14] In addition to the scooped/domed shape distinctions mentioned above, the low F0 target of the H* accent in H*HH represents a departure from the model stimuli, indicating that speakers produced non-veridical imitations of at least some of the model tunes. Note that the value for the pitch accent target in the model stimuli was the same for H* across edge tones (see Fig. 1), suggesting that this particular intonational pattern, when followed by rising F0, tends to be reproduced as a lower target. Turning to the H-L% edge tone context, similar differences between the bitonal pitch accents can be observed in Fig. 4B (tunes with H-L% edge tones), which show analogous differences in the pitch accent region as was seen in panel 4A with H-H% edge tones: the LH* trajectory has an earlier, domed rise, while the L*H trajectory has a later, scooped rise. Fig. 4C and 4D show tunes that end in L-H% and L-L% edge tones, respectively. Across panels we see the predicted edge tone distinctions: tunes ending in L-H% have a higher final F0 and end in a falling-rising contour, while tunes ending in L-L% have a lower final F0 and fall monotonically from the accentual peak. Moreover, in both panels, we see significant differences in scaling and alignment of the F0 targets among the three pitch accents. L*H has a later peak alignment compared with LH*, and an L target that is overall lower, a difference that is also observed for

[13] Successful discrimination of tunes that share a pitch accent but differ in edge tones could lead to this negative effect of preamble RMSD. For example, the pair {LH*HH, LH*LL} (top-right quadrant of the Fig. 3, near the intersection of 0.7 on the y axis and 5 on the x axis) is well discriminated as shown in Fig. 3. Conversely, tunes with different pitch accents will have different preambles, but may otherwise have perceptually confusable nuclear regions. Fig. 3 shows that the tune pair {L*HHH, LH*HH} (bottom-left quadrant of Fig. 3, with the second-lowest position on the y-axis), which differs only in the alignment of the accentual peak, is poorly discriminated, with RMSD =2 on the x-axis.

[14] The similarity of the imitations with H*HH and L*HHH is notable when compared to the trajectories that end in HL (H-L%) edge tones, shown in Fig. 4B. There we observe that the F0 target of the H* pitch accent in H*HL is clearly higher than in H*HH. Here it's important to note that the pitch accent region of the trajectories were acoustically identical in these stimuli (H*HL, H*HH); nonetheless, speakers have reproduced the pitch accent region in H*HH with a lower F0.

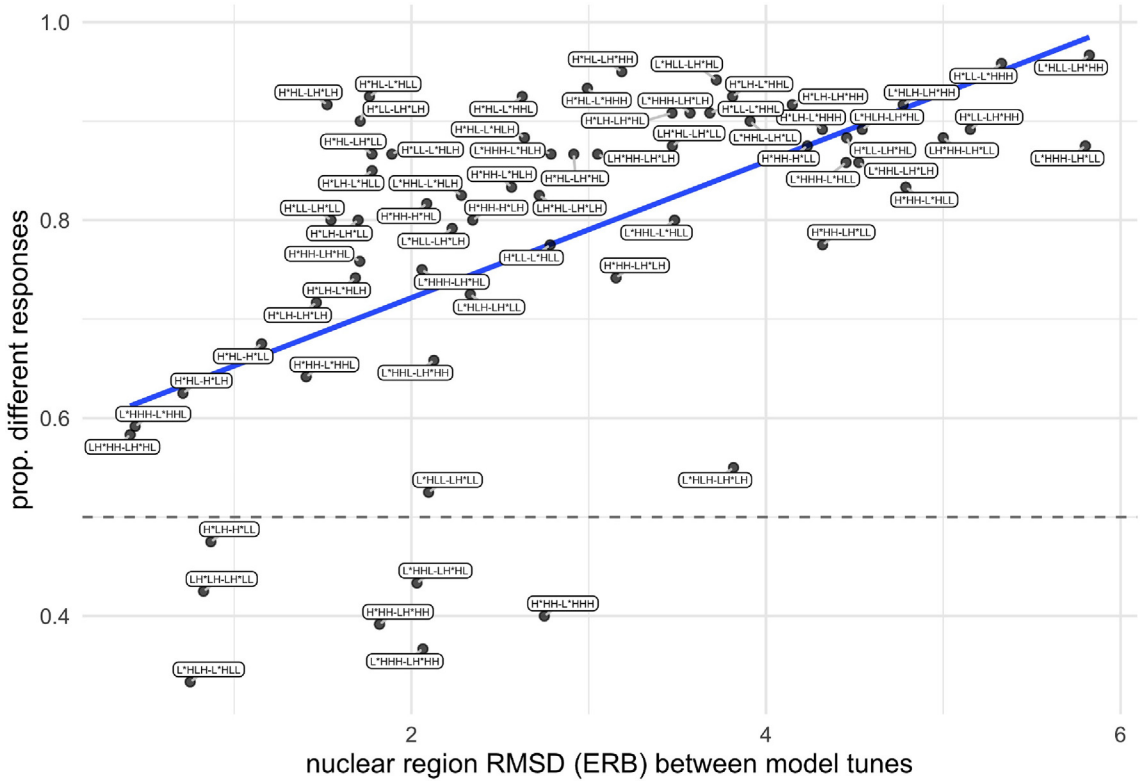

Fig. 3. Empirical perceptual discrimination responses, plotting correct "different" responses (to different-tune pairs). The x axis shows RMSD averaged by tune pair. Tune pair is labeled.

these accents by Arvaniti & Garding (2007). In the same edge tone contexts (4C, 4D), H* differs from the bitonal accents in having a lower F0 peak, with a peak alignment that is similar to LH*. All pairwise differences between tunes were assessed to be significant in some temporal interval, though some differences are small in magnitude and limited in temporal scope (see Fig. A4 in the Appendix). Given this outcome we turn now to the results of the clustering analysis, which will tell us whether the same distinctions in F0 trajectories found in the GAMM analysis emerge in the form of twelve distinct clusters in the clustering analysis.

3.3. Clustering analysis

The results from the clustering analysis are presented in Fig. 5. The optimal clustering solution returns only two clusters (Fig. 5A). Cluster 1 shows a rising-falling shape, while cluster 2 rises monotonically. The partition of the data into (only) two clusters allows us to conclude that (1) these two globally defined shapes constitute the most salient distinction in the data, and (2), that a partitioning of the data into additional clusters representing the distinctions found in the GAMM analyses is sub-optimal under the criteria of the clustering analysis. It is important to reiterate here that the two-cluster solution is not the *only* clustering solution for the data—we considered up to 12 clusters. However, it is the solution that best separated the data, as evaluated using the Caliński-Harabasz criterion, and is the one that we consider to be the best representation of the most salient distinctions present therein. Nevertheless, a two-cluster solution is very coarse grained, and begs the question of whether finer distinctions are also emergent in the data. To further probe smaller, more subtle distinctions, we carried out two additional clustering analyses for the subsets of the data defined by the first clustering analysis. In each of these subset clustering analyses, the optimal clustering solution was once again 2 clusters, leaving us with a four-way distinction that is hierarchical in nature. In Fig. 5B we identify the primary distinction as being between cluster 1 and cluster 2 and use letters a/b to refer to the subset clusters. In what follows, we consider the shape of these subset clusters and their composition in terms of exposure tunes. The supplementary materials include visualizations of two other clustering solutions. The first is the (non-optimal) clustering solution with four clusters (called "clustering-firstpass_four.jpg" in the OSF repository figures folder), which we reasoned was a good point of comparison to the four total subset clusters presented here. This clustering solution is highly similar to the one presented here and shows that our sub-setting approach arrived at essentially the same solution as a first-pass four-cluster solution. We also inspected the maximal 12 cluster solution ("clustering-12solution.jpg" in the OSF repository figures folder), as test for how well the 12 tunes mapped to these 12 clusters. That inspection led us to conclude that the mapping is noisy and imperfect, many tunes are split across multiple clusters, many clusters contain productions of multiple tunes, and no cluster contains more than 74% of a given tune production. This is consistent with our finding that many tunes cluster together in the optimal (two-cluster) solution.

First, consider clusters 1a and 1b and their mean trajectories in the top panel of Fig. 5B. The means trajectories of these

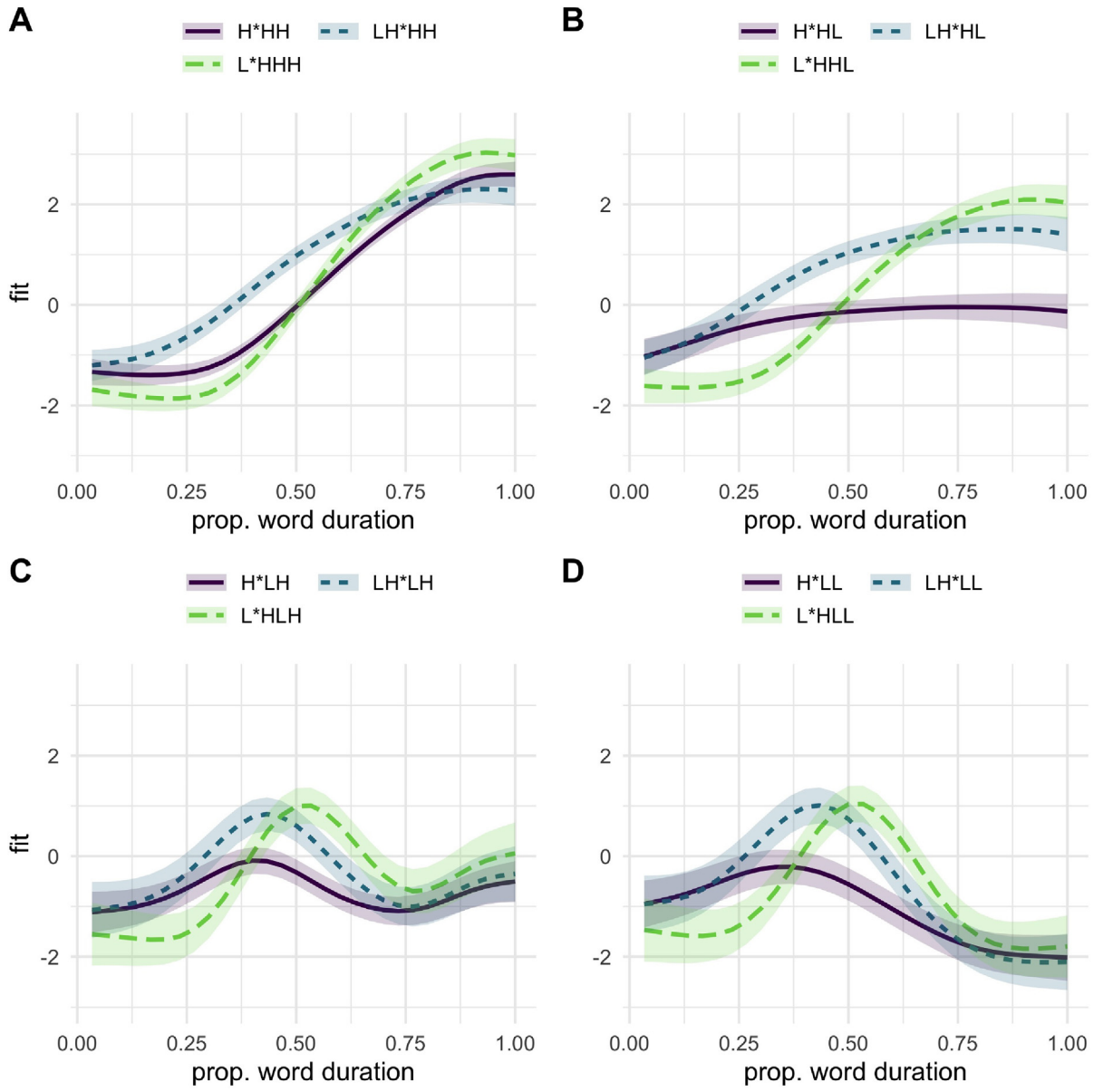

Fig. 4. GAMM fits and 95% CI for tunes, grouped by edge tone and colored by pitch accent.

two clusters are distinguished by the scaling of F0 after the initial F0 peak associated with the pitch accent. In cluster 1b, there is a fall from the initial peak to a low value, while in cluster 1a there is a much smaller F0 fall after the peak, ending in a mid-level f0. We show the relationship between the exposure tune for a given imitation (identified by the tune label of the model stimuli on the same trial) and each subset cluster in the form of heat maps at right in Fig. 5B. The heatmaps give exposure tunes in rows, and clusters in columns, with the coloration of the heatmap indicating the proportion of exposure tunes that contributed to a cluster. Note that tunes that contributed less than 5% of their trajectories to a cluster are not represented in the heatmap, so that for some tunes the total proportion (across the four panels for cluster 1a,1b,2a and 2b) does not quite add up to 1.

The relationship between exposure tunes and clusters is to some extent based on the edge tones in cluster 1a/1b. For many tunes, imitations map onto clusters based primarily on the edge tones. Cluster 1 consists mainly of imitations of tunes ending in LL (L-L%) and LH (L-H%). For instance, 97% of the imitations of tunes ending in LL are in cluster 1 with 82% in cluster 1b, while 99% of the imitations of tunes ending in LH are grouped into cluster 1 with 71% in cluster 1a. Cluster 2 has a nearly complementary composition, consisting primarily of imitations of tunes ending in HH (H-H%) and HL (H-L%). Thus, 95% of the HH tune imitations are in cluster 2 with 70% in cluster 2b. Imitations of tunes ending in HL edge tones are more widely distributed, spread between cluster 1a, 2a and 2b.

Qualitatively, we observe that the mean trajectories of the clusters 1a and 1b differ only slightly in alignment of the initial pitch movement (associated with the pitch accent), but they differ more in the scaling of the final portion of the tune, with higher (1a) or lower (1b) F0. In comparison, the F0 trajectories

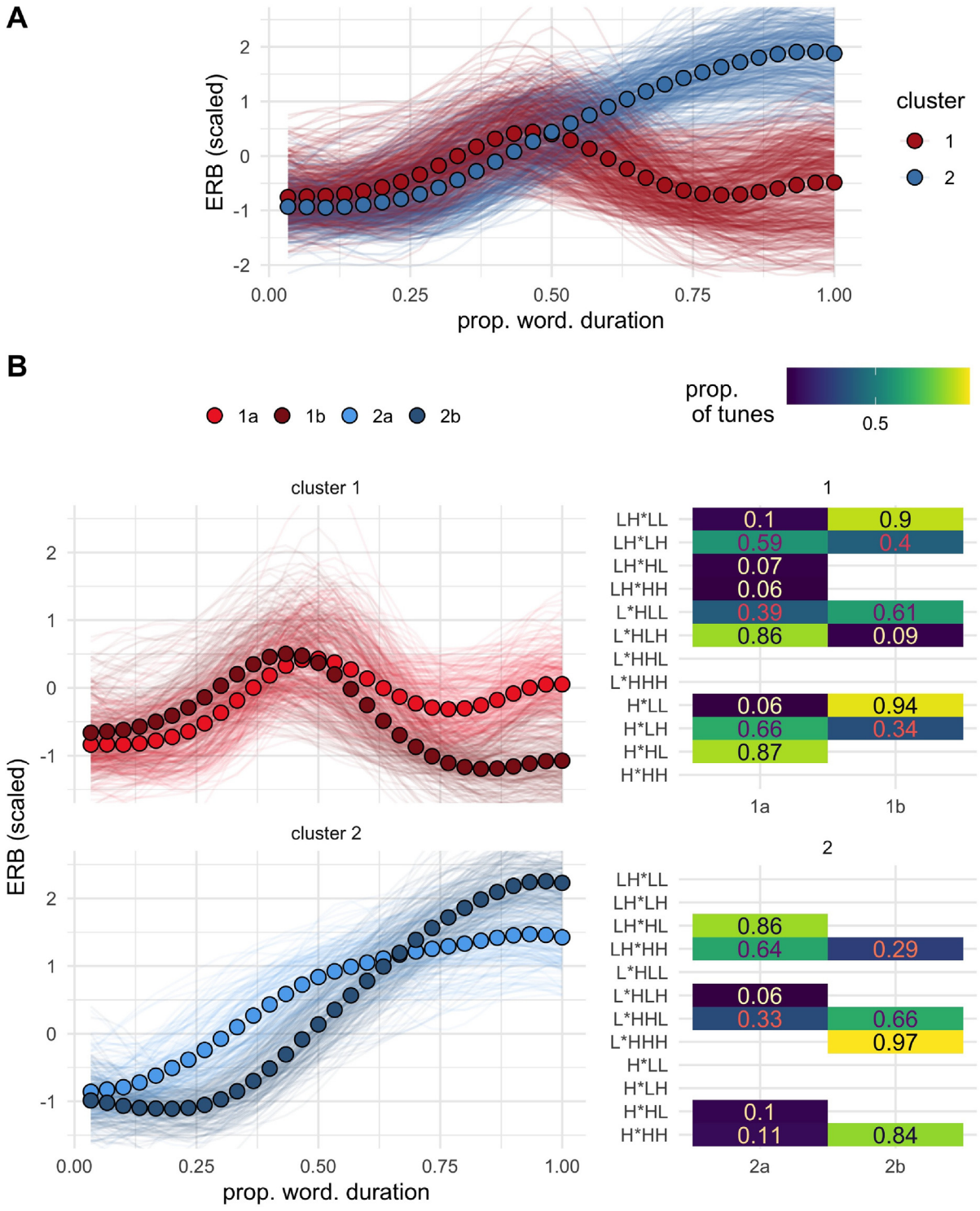

Fig. 5. Clustering results showing the first pass clustering solution (panel A) and the second pass clustering solutions (panel B). The thin lines represent mean trajectories for individual speakers, and the darker dotted lines represent the mean trajectories for each cluster. Cluster labels are given at right in panel A and at left in panel B. Heat maps at right in panel B show the proportion of each of the twelve tunes which are in each second-pass cluster (proportions under 0.05 not shown for visual simplicity).

in clusters 2a and 2b rise throughout the tune and differ in two parameters. The first is whether the rise is scooped in shape (2a) or domed (2b), and the second is the scaling of the ending F0 in each, with cluster 2a ending higher. This distinction in shape maps onto exposure tune labels: cluster 2a is composed largely of imitations of L*HHH and L*HHL, while cluster 2b is composed largely of imitations of L*HHH and H*HH, along with L*HHL. The merging of two (or more) tunes together in the same cluster indicates that variability among the imitations of either tune outweighs the variability between them, and that other distinctions in the data set are larger and more robust than those between the same-clustering tunes. The fact that the clustering algorithm finds the optimal clustering solution to be one that merges predicted tune distinctions suggests

that the merged tunes are produced with considerable variability (as these are speaker means) and are close enough to one another in F0 space to form a single cluster. The emergent clusters thus define as set of four tune shapes that are *most distinguishable* from one another, with a large collapse of many other predicted distinctions.[15]

3.4. Tonal timing across tunes

In this section we first consider the results of a model fit to the TCoG for each tune, with pitch accent and edge tone as predictors. Importantly for this analysis, we opted to consider only 9 of the 12 nuclear tunes, excluding tunes with L-H% (LH) edge tones from the analysis. As described in Section 2.4.5, the TCoG measure we employ here was developed to measure the disposition of a high F0 region in a rising-falling F0, which we extend to also test monotonic rises. It is unclear if the same model should be applied to intervals that contain multiple regions of raised F0, in our case, nuclear tunes with L-H% edge tones. For tunes ending in L-H%, a second region of higher F0 at the end of the tune will effectively pull TCoG later in time in the interval of the nuclear tune, as compared to LL, however such a holistic measure of tonal timing for multiple F0 bumps seems reductive. As such, we consider the 9 nuclear tunes in our data set which are predicted to have just a single region of high F0 (or a monotonic increase in F0).

Fig. 6A shows the mean trajectory for each of the nine tunes under consideration, grouped by edge tones. In each panel in the figure, the dashed vertical line indicates the mean TCoG for the pitch accent in that panel (in normalized time), with the line type of the dashed vertical line indicating pitch accent. In observing these means we can note some systematicities. First, in terms of edge tones, LL shows an overall earlier TCoG than HL and HH. This is expected, because for the LL shape the bulk of the F0 movement is displaced leftwards as a function of the falling movement after the pitch accent peak. In comparison, in both HL and HH edge tones, F0 rises in a manner that is essentially monotonic, displacing TCoG later in time. We also note some systematic differences based on pitch accent within each panel of Fig. 6A. In all panels, L*+H has the latest TCoG, reflecting that the bulk of high F0 for this pitch accent is displaced later in time.

We fit a model to the data to examine TCoG differences which may be present based on the predicted AM tune distinctions. The model predicted the TCoG measure described in Section 2.4.4 as a function edge tone (reference level HL), pitch accent (reference level L+H*) and their interaction. The model included random intercepts for speaker, and by-speaker slopes for the fixed effects and interactions. We also included random intercepts for the critical word in the experiment (“Helena”, “Melanie”, “Madelyn”). We step through the effects in the model to examine the extent to which there are credible differences in TCoG across each of the nuclear tunes. The model finds a credible effect of edge tones, with the TCoG for LL much earlier than the TCoG for HL and HH. Specifically, using the HL edge tones as the reference level, tunes with LL edge tones have credibly earlier TCoG alignment ($\widehat{\beta}$=−109, 95%CrI = [−117, −101], pd = 100), while tunes with HH edge tones have credibly later TCoG alignment ($\widehat{\beta}$=14, 95% CrI = [9,19], pd = 100). This difference in alignment is visible in Fig. 6B, where tunes ending in LL edge tones have a TCoG distribution that straddles the value of 0 (the boundary between the first and second syllable) and is earlier in time than the other two edge tone conditions. The difference between tunes with HH and HL edge tones, though credible, is notably much smaller than the difference between LL and either of these other edge tones (14 ms versus 109 ms).

A main effect of pitch accent was also found in the model. With L+H* as the reference level, H* showed credibly earlier alignment ($\widehat{\beta}$ = -18, 95%CrI = [-24,-12], pd = 100), while L*+H showed credibly later alignment ($\widehat{\beta}$ = 35, 95%CrI = [28,-42], pd = 100). This pattern is visible in Fig. 6C, which shows TCoG for different tunes split by both edge tones (facets in the plot) and pitch accent (rows). This relationship between TCoG and pitch accent, with H* earlier than L+H* earlier than L*+H, is clear in both the L-L% (LL) and H-L% (HL) facets in Fig. 6C. Notably however, the H-H% (HH) panel deviates from this pattern, whereby H*HH shows a mean TCoG that is later in time than L+H*HH. This can be explained if we look to the productions of H*HH and L+H*HH from the GAMM fits in Fig. 4, and trajectory means in Fig. 6A. In the context of the HH edge tones, the H* trajectory more closely resembles L*+H than L+H*.[16]

Pairwise comparisons of all pitch accent and edge tone combinations, representing each of the nine tunes in this model, were also carried out using *emmeans* (Lenth, 2021). The goal here was to see if credible differences in TCoG could be found for all pairwise comparisons between tunes. Strong evidence for pairwise differences (pd = 100) was found for all but three pairs of tunes. Only moderate evidence (pd = 95) was found for a difference between H*LL and LH*LL (see the small difference in the LL panel in Fig. 6A, and 6C), and between H*HH and L*HHH (pd = 97, 95%CrI just narrowly excluding zero). One pair did not show reliable evidence for a difference in TCoG: H*HH versus L*HHL (pd = 92). In summary, the TCoG model with pitch accent and edge tone labels shows that both of these tune components influence tonal timing as measured with TCoG, with credible pairwise differences between most tunes. However, some of those TCoG differences are very small, with heavily overlapping distributions as shown Fig. 6B and 6C.

We next tested the extent to which the emergent clusters were distinguished by TCoG. Fig. 7A shows the mean trajectories for each cluster (excluding LH edge tones), with the mean temporal TCoG in normalized time plotted as a dashed vertical line. Fig. 7B shows density plots for the distribution of TCoG across emergent clusters. We ran a second model of TCoG, this time predicting TCoG variation as a function of cluster

[15] The online supplementary materials contain a description of an alternative approach to clustering, for which we used tonal center of gravity parameters, temporal TCoG (described above) and “frequency TCoG” (described in the supplement) and clustered over the tunes in the two-dimensional parameter space using traditional k-means clustering. Though the parameterizations of the tunes are quite different, the two analyses largely agree on which tunes cluster together. Details on this alternative method, and brief discussion of similarities to the time-series k-means clustering analysis are discussed in the online supplement.

[16] The H*HH trajectory has a rise shape that is more similar to L*+HHH than to L+H*HH. It is this difference that we see reflected in the later TCoG of H*HH, which corresponds to a later, and scooped rise shape. This is also reflected in the model in a credible interaction between the HH edge tone and H* pitch accent ($\widehat{\beta}$ = 45, 95%CrI= [36, 53], pd = 100), indicating a later TCoG for this combination than reflected in the main effects.

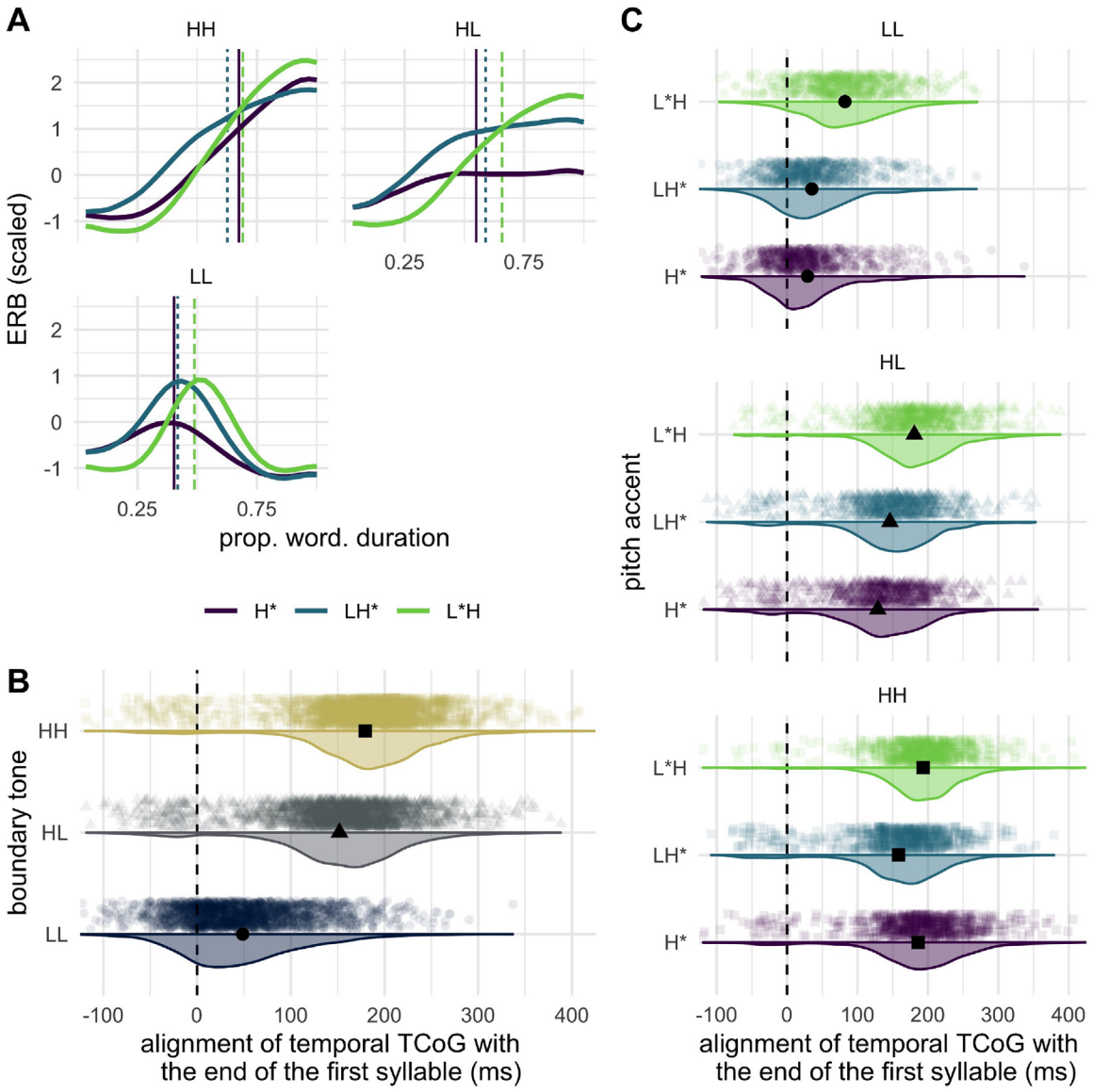

Fig. 6. Trajectories for 9 tunes with the mean temporal TCoG in normalized time indicated as a vertical line, matched in color to the corresponding pitch accent (panel A), and TCoG, as aligned with respect to the first syllable, split by edge tones (panel B) and by edges tone and pitch accent (panel C). The value of zero marks the end of the first syllable in the nuclear word.

label with four levels (1a, 1b, 2a, 2b), with otherwise the same structure as the model that was fit to edge tone and pitch accent categories (random intercepts for speaker and critical word, with by-speaker slopes for cluster). A comparison between clusters finds that there was a credibly earlier TCoG in cluster 1b versus cluster 1a ($\widehat{\beta}$ = 78, 95%CrI = [72, 84], pd = 100). As shown by the cluster mean shapes, this results from a more scooped rise in cluster 1a, with higher F0 at the end of the tune, both of which displace TCoG later in time. There was additionally a credible difference between cluster 2a and 2b, with cluster 2b showing credibly later TCoG ($\widehat{\beta}$ = -36, 95%CrI = [-40, –32], pd = 100). As shown in Fig. 7A, this difference is the result of a more scooped rise and higher ending F0 in cluster 2b, both of which displace TCoG later in time. We take this latter result to suggest TCoG is an appropriate metric for quantifying shape-based distinctions in monotonically rising tunes, which is a new application of the measure to our knowledge. There were additionally credible differences between all possible comparisons among the four clusters including across clusters 1 and 2, with all pd = 100. These models show that TCoG is a parameter that effectively separates the emergent clusters and captures shape-based distinctions that are observable in the clusters.

To further test the predictive power of TCoG in explaining the mapping of imitative productions to a given cluster, we modeled TCoG as a function of cluster, but this time *within* tune category, that is, modeling the effect of cluster membership on TCoG for a given tune. As described in Section 3.3, there is some variation in how the production of a given tune maps to a given cluster, and the models within a given tune effectively test the extent to which TCoG predicts this variability. In Fig. 7C the TCoG measure is shown for each of the 9 tunes we tested, with each panel split by cluster. As an example, note that L*HLL (bottom right corner) contributed to both cluster 1a

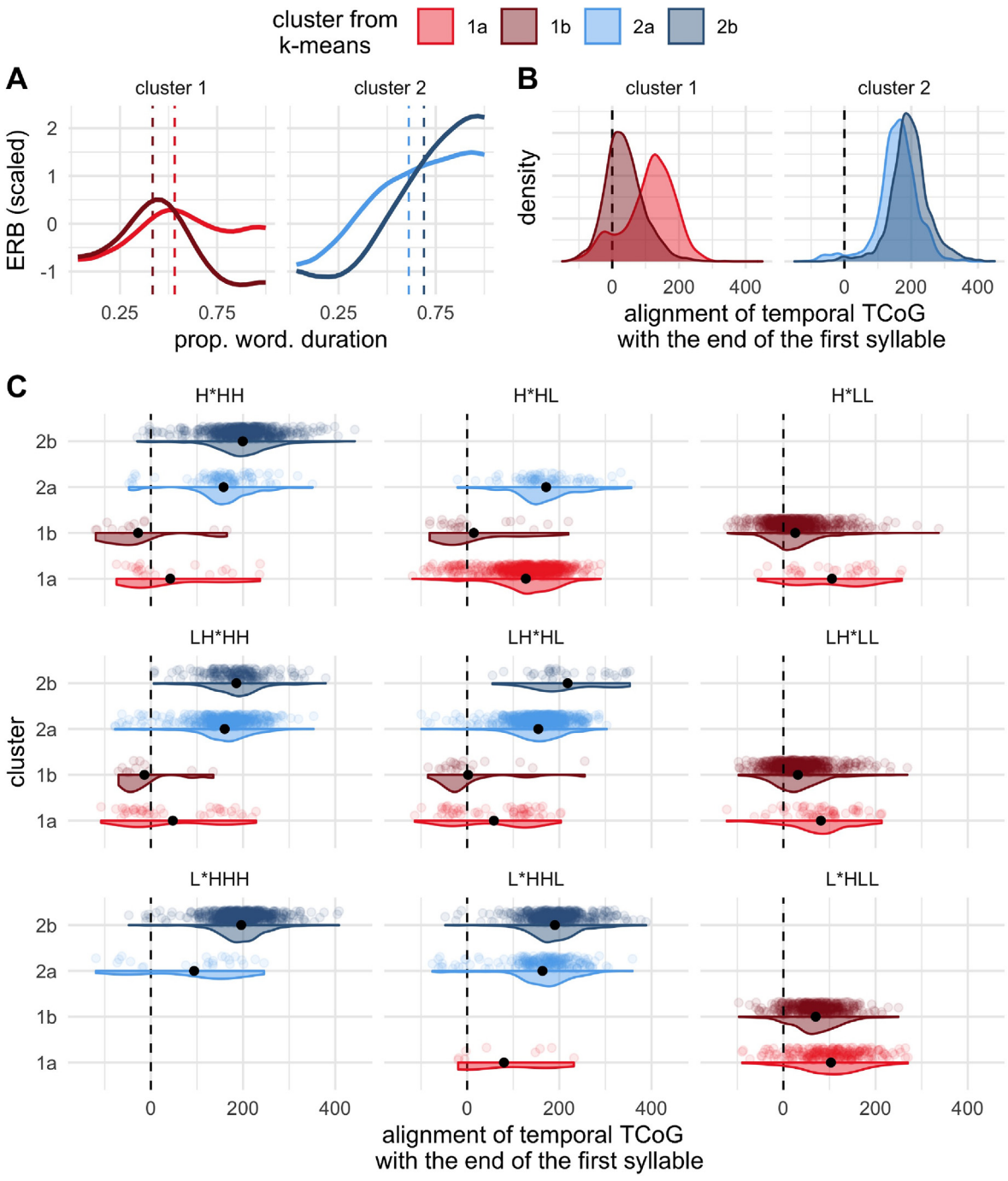

Fig. 7. TCoG as a function of the cluster labels from k-means clustering. Panel A shows the trajectory means by cluster (excluding LH tunes). Panel B shows the distribution of TCoG values for each cluster, panel C shows TCoG values for each tune, split by cluster.

and 1b. Productions of L*HLL with overall later TCoG are grouped in cluster 1a, in line with cluster 1a having later TCoG than cluster 1b.

Table 3 summarizes the differences in TCoG across clusters within a given tune, mirroring the data that is in Fig. 7C. The full model summaries are available on the open access repository. As shown in the Table, there are credible differences in TCoG across clusters, for every tune. For example, LH*HL contributes a fair number of tokens to both cluster 2a and cluster 2b, and the TCoG model for that tune finds that

Table 3
Summary of cluster effects on TCoG within an exposure tune, organized with the same layout as Fig. 7C. When < separates two cluster labels this means there was evidence that they differ in TCoG, with pd values given in parentheses.

	HH	HL	LL
H*	{1a,1b} < 2a < 2b (pds > 98)	1b < 1a < 2a (pds = 95)	1b < 1a (pd = 100)
LH*	{1a,1b} < 2a < 2b (pds > 98)	1b < 1a < {2a,2b} (pds > 98)	1b < 1a (pd = 100)
L*H	2a < 2b (pd = 98)	{1a, 2a} < 2b (pds > 98)	1b < 1a (pd = 100)

productions that fall into cluster 2b have later TCoG. Analogous effects can be seen for all tunes, whereby the division of a tune's productions into emergent clusters predicts TCoG. What these analyses show together is that variability in a tune mapping to a cluster can be explained by aspects of tonal timing for that particular production of a tune. In other words, when a single tune maps to multiple clusters, it is due to variation in that tune that is captured by TCoG. We take this to suggest that the clustering partition not only captures the most salient distinctions in the data, it also captures subtler within-tune variations in tonal timing.

3.5. Perception analysis with cluster labels

The speech production data presented up this to point have shown subtle distinctions among all 12 tunes we tested (GAMMs and TCoG models), while the bottom-up clustering approach defined a cluster partition with many fewer distinctions: a principal distinction based on whether tunes rise monotonically to a high value or fall from an accentual F0 peak. Secondary distinctions determined by the clustering algorithm were also reflected in the models of TCoG variation for each tune, which captured the displacement and shape of the monotonic rise (for the rising cluster 2), and alignment and extent of fall for the other cluster (cluster 1). We take the clustering results to reflect the primary distinctions among tunes elicited in this experiment, and we hypothesize that these distinctions are also the most perceptually salient. As such, we predict that the clustering partition of the data may effectively predict variation in listeners' perceptual discrimination of tunes. We tested this prediction by evaluating additional models of the perceptual discrimination responses that include clustering information in two forms. In the first, we predicted perceptual discrimination responses by RMSD (as in the model described in Section 3.1) and an additional variable that defines if tune pairs clustered together or not. We call this "cluster class". This variable may be "within" for two tunes that were grouped together in the first clustering analysis (Fig. 5A), e.g., H*HH and L*HHH. The variable is otherwise "between", i.e., for tune pairs that were grouped into separate clusters, e.g., H*HH versus H*LL. Another model was fit which further differentiated the "within" cluster variable, to distinguish tune pairs that are grouped together within the rising cluster (cluster 2 in Fig. 5A), from pairs grouped together within the non-rising, cluster (cluster 1 in Fig. 5A). Splitting the "within" variable apart in this way allows for the possibility that perceptual discrimination performance may vary not only by whether a tune pair is partitioned between or grouped together within the two primary clusters, but also for the possibility that variation within these two emergent cluster classes may be discriminated to a greater or lesser degree. We refer to this model as containing ternary clustering information. We carried out model comparison to assess if these added variables led to an improved fit of the model using the *loo* package (Vehtari, Gelman & Gabry, 2017; Vehtari, Gabry, Magnusson, Yao, Bürkner, Paananen & Gelman, 2020). This package computes leave-one-out cross validation (LOO-CV) to estimate model prediction accuracy. We compare this measure for three models: the model fit with only RMSD, the model fit with RMSD and the binary within/between cluster classification, and the model fit with RMSD and the ternary division of between cluster, within-rising and within-non-rising. The best model was the one with ternary clustering information, suggesting a possible difference between within-rising and within-non-rising cluster classes, i.e. one class tends to be discriminated more accurately than the other. (note that the binary clustering model was also better than the model with no clustering information). The online supplementary materials contain information about this model comparison, as well as models that consider two additional ways of computing differences between the stimuli (instead of RMSD). One of these, suggested by a reviewer, was TCoG, and the other was computed by combining TCoG and RMSD information. The model that is discussed below is the one which was determined to have the best fit of all of these using LOO-CV procedures (see the online supplement for details).

The best fit model also showed a credible influence of RMSD (in the nuclear region), such that higher RMSD led to improved discrimination of tune pairs ($\widehat{\beta}$ = 0.45, 95% CrI = [0.30, 0.61], pd = 100), but there was an additional credible influence of the three-level cluster class variable. Taking the "within-non-rising" set as the reference level, it was observed that "within-rising" tune pairs showed an overall lower proportion of different responses ($\widehat{\beta}$ = -1.19, 95%CrI = [-1.49, −0.90], pd = 100), in other words, tune pairs from the within-rising class (cluster 2 in Fig. 5A) were harder to discriminate than tune pairs from the within-non-rising class (cluster a in Fig. 5A). Additionally, between-cluster tune pairs show a credibly higher proportion of different responses as compared to within-non-rising pairs ($\widehat{\beta}$ = 1.20, 95%CrI = [0.74, 1.68], pd = 100). We thus have evidence that both RMSD and cluster class help predict discrimination responses, with cluster class patterning as between > within-non-rising > within-rising. These effects are visible in Fig. 8A, which plots the empirical data, and Fig. 8B, which plots the model estimates for the main effects of cluster class. In Fig. 8A, for which each point represents a tune pair for a given stimulus item (model speaker and model sentence combination), we see the aforementioned distinction in cluster class in term of the overall vertical position of the regression lines over RMSD. The effect estimates in Fig. 8B allow us to examine this distinction as a main effect, showing that within-rising perceptual discrimination is overall at or below chance. Within-non-rising discrimination is higher but the 95% CrI still narrowly includes chance (0.50). Between-cluster discrimination is much higher, and well above chance.

There was additionally a credible interaction in the model between cluster class and RMSD: greater RMSD adds little to perceptual discrimination for the within-rising cluster class, comparing to the reference level of within-non-rising ($\widehat{\beta}$ = -0.31, 95%CrI = [-0.50, −0.11], pd = 100). There was no evidence for an interaction with the between-cluster class and the reference level of within-non-rising (pd = 53). To examine the interaction in more detail we extracted the estimated effect for RMSD in each cluster condition using the *estimate_slopes ()* function from the *modelbased* package (Makowski, Ben-Shachar, Patil & Lüdecke, 2020). This function extracts estimates for the effect of RMSD, showing a similar effect for both the within-non-rising cluster class ($\widehat{\beta}$ = 0.45, 95%CrI = [0.30, 0.61]) and between cluster class ($\widehat{\beta}$ = 0.46, 95%CrI = [0.27,

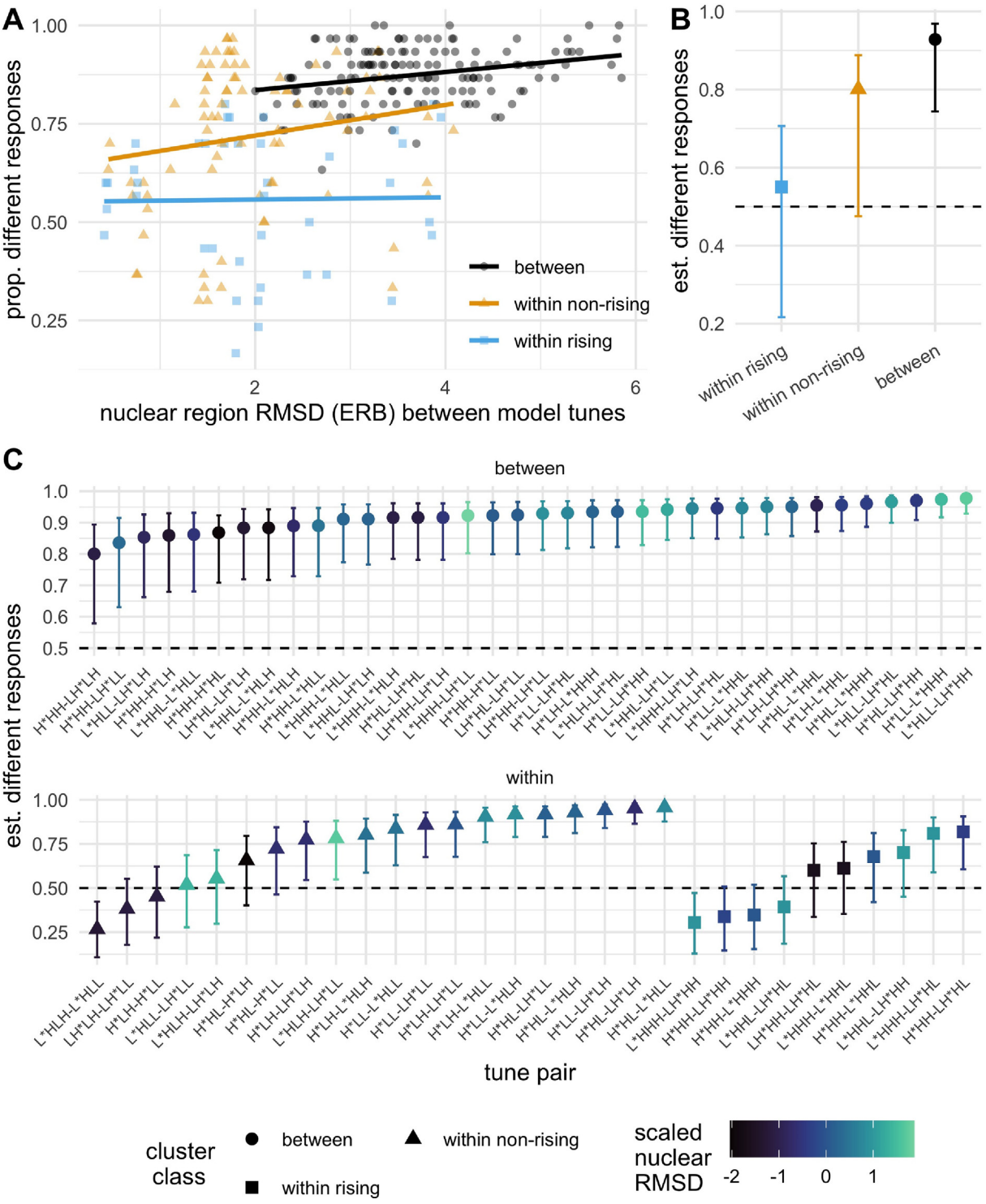

Fig. 8. Panel A: Means for empirical "different" responses as a function of RMSD, with points and linear regression lines color coded by cluster class. Panel B: Model estimates and 95% CrI for cluster class, showing the main effects for this variable from the cluster and RMSD model. Panel C: Model estimates and 95% CrI for tune pair from the tune pair model, sorted by cluster class and colored by RMSD. Note that the y axis ranges in the two rows are different. The horizontal dashed line indicates 50% (chance).

0.68]). In comparison, there is *not* a credible effect for the within-rising class, with credible intervals for the estimate (narrowly) including zero ($\widehat{\beta}$ = 0.15, 95%CrI = [-0.02, 0.32], pd = 96). There is thus only weak evidence for an effect of RMSD in the discrimination of tune pairs in the within-rising class: in other words, irrespective of measurable differences in RMSD, tunes with monotonically rising shapes are poorly discriminated from one another. This interaction can be seen in the empirical data in Fig. 8A in the generally upwards sloping lines for between and within-non-rising classes left to right along the x axis, with the line for within-rising tune pairs remaining relatively flat.

We ran one additional model, as a test of how specific tune pairs are discriminated. This model predicted perceptual dis-

crimination responses as a function of tune pair as the sole predictor: a categorical factor coding 66 possible pairs. Random effects in the model were random intercepts for speaker and base item (model speaker and sentence, with four levels). The goal of this analysis was to identify specific tune pairs that were discriminated at or below chance (those for which the model predicted 95%CrI does not exclude chance, i.e., 0.50 estimated "different" responses). These model predictions are shown in Fig. 8C, where tune pairs are sorted by cluster class, and ordered by the proportion of estimated different responses. The coloration of the points additionally shows RMSD which, importantly is scaled *within* each cluster class, showing relative RMSD difference in each class. First consider just the between cluster tune pairs, shown in the first row of Fig. 8C. All are discriminated well above chance, confirming the model with RMSD as a predictor in showing that any tune pairs that are grouped into different cluster classes are successfully discriminated by listeners. Examining tune pairs in the within-non-rising cluster class (bottom row, right Fig. 8C) allows us to see which tune characteristics lead to poor discrimination, from which we can extract several generalizations. First, we observe that the three tune pairs in the within-non-rising class that have the lowest discrimination accuracy consist of tunes that have the same pitch accent but differ in their edge tone (LH vs. LL). The RMSD difference between tunes in each of these three pairs is relatively small, as indicated by coloration, and discrimination performance is at or below chance. Clearly, this boundary tone distinction is not well perceived by listeners, which is also reflected in the clustering analysis of imitative productions). Next, we observe two tune pairs in the within-rising class that are also discriminated below chance, yet which have higher RMSD (brighter coloration). They are {LH*LL vs L*HLL} and {LH*LH vs L*HLH}; the tunes in each of these pairs vary in pitch accent alignment and share the same edge tones. Poor discrimination of alignment differences suggests this is a parameter that by itself is insufficient for perceptual discrimination. Two additional tune pairs in the within-rising cluster class are also discriminated at chance; both pairs include H*HL, which is compared to H*LL or H*LH. These pairs are low in RMSD as well, and offer some further evidence that edge tone distinctions, particularly in a small pitch range as with these three tunes, can be hard to perceive.

Finally, consider the within-rising class shown at right in the second row of Fig. 8C with square symbols as points. All but two tune pairs in this cluster class are perceived at or below chance (aligning with the RMSD-based model). Many of these pairs also differ principally in the alignment of the accentual rise, for example {LH*HH vs L*HHH} and {LH*HL vs L*HHL}. H* is also highly confusable with the other two pitch accents in the HH edge tones context. As in the other cluster class, these distinctions in alignment can lead to large differences in RMSD, which notably do not carry over into successful discrimination performance. Only two pairs within this class show discrimination above chance (though still far from ceiling), these two pairs varying in both pitch accent and edge tones.

In summary, the addition of clustering information to the perception data improves model fit, indicating that cluster classes show clear predictive power in addition to RMSD values, with the interaction seen in the model further indicating that the impact of RMSD is mediated by cluster class (within-rising tunes being discriminated poorly regardless of RMSD). Further, in testing which particular tune pairs are discriminated at or below chance we find several commonalities in the poorly discriminated pairs: differences in edge tone only, which tend to have small RMSD differences are poorly discriminated, and differences in accentual peak alignment only (especially in monotonically rising tunes), which can lead to large RMSD differences, are also generally poorly discriminated.

4. Discussion

The present study examined the production and perception of 12 nuclear tunes in American English, which were defined on the basis of pitch accent, phrase accent and boundary tone labels. We focused on three pitch accents, H*, L+H*, L*+H, for which past research has questioned the existence of three discrete categories. Using an imitative speech production study with 70 speakers, we assessed distinctions among these tunes using GAMMs, bottom-up k-means clustering analysis, and analysis of tonal timing via Tonal Center of Gravity. We also addressed the perceptual discriminability of idealized model tunes using an AX perception task (with 60 listeners).

To synthesize the foregoing results, two core questions from the introduction are revisited. First, is the AM model correct in predicting 12 nuclear tunes which behave as categories? Second, do the standard descriptions of the phonetic (F0-based) differences characterize our data? Evidence for a phonological category could be separability in intonational form (in at least some contexts), which maps to a difference in intonational meaning. We are not in a position to address the meaning-based component of this question, as our study lacks examination of possible tune meaning, and the field at large lacks a solid empirical base to begin answering this question, as described in Section 1. However, with the present data and multi-faceted analyses we present above, we are well-positioned to address the question of intonational form. That is, what intonational forms are both reliably perceived as distinct and produced as distinct by speakers and listeners? First, consider the perception data. We assume that two phonologically distinct tunes should be perceived as different with reasonable consistency, allowing for listener variation and noise. If we take this as a heuristic, the perceptual discrimination data over tune pairs (Fig. 8C) indicates that a substantial number of tune pairs are discriminated at or below chance, suggesting a lack of categorical distinction. Crucially, these patterns relate to the emergent "cluster classes" within the production data as described in Section 3.5. Two tunes that are partitioned into different top-level clusters (cluster 1 vs. 2), are successfully discriminated above chance in all cases. In this sense, we take the primary emergent clusters to define a distinction among tunes that is perceptually very robust, namely, whether or not the tune contains a monotonic rise in F0 to a high value. The perception results also provide us with a key insight about perceptual discriminability for tunes that are grouped together in the emergent clusters. Almost all tune pairs from the rising cluster are discriminated at or below chance (cf. Dilley & Heffner, 2013), while tune pairs from the non-rising cluster show more variability and their perceptual discrimination is predicted well by RMSD (unlike for rising tunes). There are some notable differences in tune discrimination as a function of edge

tones that illustrate this distinction. For example, H*LL and LH*LL are discriminated above chance in Fig. 8C, while H*HH and LH*HH are discriminated below chance (the stimuli for these two pairs of tunes notably are acoustically identical in the pitch accent region). If we take the emergent clusters and perceptual discrimination data together, we can conclude that monotonically rising tunes are (1) robustly differentiated from other tunes in all cases, (2) poorly differentiated from each other and (3) define an emergent class from the analysis of unlabeled data. These three points together lead us to conclude that the (monotonically) rising tunes together evidence category-like behavior. The emergent rising versus non-rising partition of the data can in this sense be seen as a primary, and categorical distinction among these nuclear tunes, as realized with F0.

The perceptual data for non-rising tunes is more complex, with some distinctions discriminated above chance. Most notably, two pairs of tunes with large RMSD and alignment-based differences are perceived as the same by listeners: L*HLH versus LH*LH, and L*HLL versus LH*LL. The perception data thus show that, when edge tones are held constant, the distinction between LH* and L*H accents is perceptually tenuous (when presented along with many other tune pairs as in our experiment). While we take below-chance discrimination performance to evidence a lack of category-like distinctiveness, we should be cautious in assuming that above-chance performance entails a necessary categorical distinction, particularly in light of an alternative continuum-based model for peak alignment and peak scaling (Ladd & Schepman, 2003). That is, successful discrimination of a tune pair could be the result of phonetically salient variation of a parameter that varies along a graded/continuous dimension. For example, the continuum-based account of H*(LL) and L+H*(LL) (e.g., Ladd & Schepman, 2003, Ladd, 2022) would hold that these tunes are well discriminated as they represent good exemplars of endpoints of a continuous variation within a single category. Herein, the clustering analysis can offer insight into the extent to which these distinctions are emergent. The clustering analysis for the non-rising cluster partitions subcluster 1a and 1b principally on the basis of edge tone, not pitch accent, showing that in the face of varying edge tone contexts, the most salient (emergent) distinctions in the data are those that are based on edge tone, not pitch accent. This reinforces the idea that pitch-accent based distinctions among the three pitch accents tested in this study (H*, L+H*, L*+H), across our 70 speakers, are not systematic enough to be discovered in bottom-up clustering for tunes in the non-rising cluster.

Considering only the perception data and clustering results then, we might be inclined to conclude that many distinctions proposed by the AM model are spurious. However, according to the GAMM and TCoG models, in which imitated tunes are associated with the ToBI labels of their corresponding stimuli, this is not the case. Each of the 12 tunes was found to be distinct from the others in the GAMM analysis, and to differ in tonal timing (though sometimes by quite small measures) according to the TCoG modeling. Considering all analyses together, we find that some tune pairs are not reliably perceived as different, do not emerge as distinct in clustering, and yet, are separated by (small) differences in F0 shape and timing. Here it is important to note that AX discrimination is a metalinguistic judgment that may not directly reflect what is perceived in the auditory signal. That is, the listener may perceive a subtle distinction in F0 between two auditory models yet judge that distinction to be below a threshold that would generate a “different” response. One case where we would expect this sort of mismatch in production and perception is if a listener perceives an F0 difference as within- rather than between-category variation for the tune category distinction invoked in making a same/different judgment. If we allow for the representation of within-category variation in encoding and as a production target, this detail may be evident in imitative productions, in the spirit of a continuum-based distinction among AM categories as in Ladd & Schepman (2003) and Gussenhoven (1984). In other words, a participant may hear phonetic detail that is reflected in their imitations, but not used for AX same/different judgements.

4.1. Rising nuclear tunes: Hierarchical and context-dependent distinctions

How should the present data inform a theory of intonational phonology? Most fundamentally the data suggest that distinctions among nuclear tunes are *hierarchical.* The distinction between rising and non-rising tunes is a fundamental one, with finer grained variation evident within the emergent rising and non-rising classes. In this sense, the AM model of a nuclear tune as the unconstrained combination of pitch accents, phrase accents, and boundary tones misses the fact that some of these combinations are perceived to be the same by listeners judging our stimuli, and evidence only subtle differences in imitative productions. Other combinations of pitch accent, phrase accent, and boundary tone are nearly always perceived as distinct and define separate emergent clusters, an asymmetry that we believe is desirable for a model to capture. The data thus speak to the possibility that the 12 nuclear tunes should not be viewed as a set of intonational forms of equal status, but rather as a structured set. Some distinctions are primitive, categorical, and robust, while others are secondary, gradient, and variable.

The principal distinction, discussed above, boils down to the phrase accent label in the AM model, reflecting whether a tune rises after the pitch accent or remains high (H-), or falls from the accentual peak (L-). Below this highest-level distinction in the hierarchy, how should we consider the next level of distinctions evident among tunes? Given the nature of the tunes we investigate here, it is clear that tonal timing plays an important role. Within each of the emergent clusters, we observed variation in TCoG showing that clusters differed substantially in their timing of TCoG, and furthermore that *within-cluster* variation, i.e., variation in how imitations of a particular tune map to a particular cluster, is also predicted by TCoG. For example, imitations of L*HHL are split between cluster 2a and 2b, and this split is reflected in TCoG in the sense that productions of L*HHL which fall into cluster 2b have later TCoG. The clustering distinctions that are emergent for the rising clusters thus subsume variation in TCoG both *across* tunes and *within* tunes, which suggests that the two rising sub-clusters 2a/2b span a continuum that can be captured with TCoG as a measure of tonal timing. If we take the cluster means as approximate representations of the endpoints of this continuum, at

one end there is a rising shape with a domed rise and flatter ending F0 (more like HL), while at the other there is a scooped rise with higher ending F0 (more like HH). The co-occurrence of a domed rise with lower ending F0 and a scooped rise with higher ending F0 in monotonically rising F0 movements jointly create distinctions in temporal TCoG: a domed rise and lower ending F0 pulls TCoG earlier in time, a scooped rise and higher ending F0 pushes it later. It appears that this pattern of co-variation between rise shape and ending F0 co-occurrence is not accidental, but rather represents a systematic distinction in tonal timing that varies along a continuum between clusters 2a and 2b. Also considering that perceptual discrimination among most tunes in the rising cluster is not above chance, we suggest that the group of rising tunes may be considered a single category, with potentially meaningful variation in tonal timing within that category (notably, the only two rising tune pairs that were discriminated above chance paired LH*HL – a domed early rise –with two tunes that show a scooped/late rise, L*HHH and H*HH).

Though the results suggest a continuum of tonal timing between emergent rising clusters 2a and 2b, they leave unsettled the question of whether this represents a continuum in which intermediate steps are each equally probable, or an attractor space, in which particular regions along the continuum are more densely populated and generate category-like structure (Roessig, Mücke, & Grice, 2019). An attractor space is conceptually commensurate with the view discussed by Gussenhoven (1984), in which a gradient dimension has "preferred" positions. In comparison to the rising tunes, the set of non-rising tunes is more heterogenous, and shows variation in the ending F0 shape, the scaling of the F0 peak, and alignment of that peak. As noted above, the emergent clustering distinctions in subclusters 1a/1b do not capture the predicted pitch accent distinctions, because H*, L+H* and L*+H cluster together. The observation of predicted scaling distinctions (in the GAMM) and timing distinctions (in GAMM and in TCoG modeling) nevertheless shows that these pitch accents are differentiated from each other, to some extent, in the imitated productions. However, the magnitude and consistency of those differences are not substantial enough to generate distinct clusters. This, like the results from the rising cluster, is suggestive of (potentially meaningful) variation within emergent classes in F0 scaling and alignment.

Our findings also crucially reveal *context dependence* in how a distinction is perceived and produced. In the perception results, it is clear that perception of distinctions between pitch accents depends crucially on the edge tone context, with rising edge tones eliminating perceived distinctions as described in Section 4. (cf. Dilley & Heffner, 2013, who found fairly continuous variation in tonal timing for imitations in rising contexts). A potential implication for intonation theory is that intonational tunes are best viewed not as series of independent phonological units that license distinctions in every combination, but rather as context dependent F0 targets. The GAMM modeling reinforces this idea, in showing that the production of F0 for a particular pitch accent, e.g. H*, is fundamentally different as a function of edge tone context (comparing across panels in Fig. 4). The H* pitch accent in H*HH is very similar to the L*H pitch accent in the L*HHH tune, and to the LH* pitch accent in LH*HL, suggesting the lack of robust contrast in this context, and in line with the emergence of a rising cluster class that contains these tunes.

The issue of context dependence can also be related to the question of "tone phonotactics", and the observation that tonal combinations (as represented with ToBI labels) vary substantially in their frequency (Dainora, 2001, 2006). Taking the example above, the lack of a distinction between tunes could be described as "neutralization" of, for example, H* and L*H in the context of HH edge tones. The extent to which this should be considered neutralization of two underlying phonological categories as compared to a single category which exists in this edge tone context, remains to be determined. However, given that listeners were exposed to F0 patterns with substantially different F0 trajectories for e.g., {H*HH, L*HHH} and yet failed to produce or perceive them as distinct suggests to us that there may not be a category distinction to be neutralized in the first place. Positing contextual neutralization for tones in this sense commits to the notion that two pitch accents (for example) are discretely different (in a particular context), which is not strongly supported by the present data.

One possibility to test this idea further would be to vary the context in which tunes are produced to see if there is any context where the proposed distinction is robust (discussed more below). If there are tonal contexts (e.g., prenuclear tune contexts), or metrical contexts, in which distinctions of the sort referenced above emerge, this could clearly be taken as evidence in favor of the neutralization account. This line of work examining context and neutralization should further consider how they relate to the frequency with which tunes are produced, in light of the evidence for probabilistic tone phonotactics shown in Dainora (2001, 2006). Poorly-differentiated tunes in our data set may be tunes that are infrequent or have a high-frequency similar tune (e.g., LH*LL and L*HLL, according to Dainora, 2006). However, here we feel that direct comparison to Dainora's (2001, 2006) probabilistic model is difficult. There are for example, very few rising tunes in that corpus, possibly due to the radio news speech style or the small number of speakers (two). Further, that model presumes discrete tone units, which is an assumption the present study does not make. Nevertheless, we think future work should crucially consider frequency of use (within particular discourse contexts, and across contexts), as a predictor for how well tunes are differentiated from one another, building on the insights from Dainora (2001, 2006).

Finally, a comparison to the findings in Cole, Steffman, Shattuck-Hufnagel, & Tilsen, 2023 is warranted. As noted in the introduction, Cole et al. tested the imitative productions of eight nuclear tunes formed by the combination of H* and L* pitch accents with all edge tone configurations. They found five emergent clusters in an analogous time-series clustering analysis to the one presented here. Those emergent clusters essentially collapsed tunes that had the same pitch accents (H* or L*), but which varied in edge tone configuration such that the distinction between H* and L* was well-preserved. This presents a clear departure from the present results, in which clustering partition of the data is based primarily on edge tones in the first pass analysis, and on pitch accent only for the rising cluster tunes in the second round of clustering. Though timing differences were indeed measurable in imitative productions between pitch accents, they were overall small, and

poorly discriminated in perception (especially for rising-cluster tune pairs). The comparison of our results to Cole et al. thus clearly suggests that the *type of distinction* between H*, LH* and L*H can be understood as different from the distinction between H* and L*. In other words, in line with the preceding body of work discussed in the introduction, the evidence for H* and LH* as robustly distinct pitch accents is not strong, nor is the evidence for LH* versus L*H. Conversely, the distinction between H* and L*, as examined by Cole et al., is quite robust. Looking across studies then, we have evidence for another sort of hierarchy: a primary, or categorical, distinction between high/rising pitch accents and low (L*) pitch accents, with secondary, or graded, distinctions within the high/rising category. In the extreme, this suggests a model of intonational phonology with a categorical high/low phonological distinction, and with graded variation in the high category resulting in H*, LH* and L*H -like shapes. Ladd (2022) evokes a similar possibility in his discussion of H* versus LH* in particular: "The extent (and probably the timing) of the rise in pitch preceding the accentual high are manipulated by the speaker to express a variety of nuances, but the basic phonological choice is in all cases 'high accent'" (p 253). Though meaning-focused research is needed to further this line of work, this strikes us as a promising future direction to probe how {H*, LH*, L*H} distinctions are different from, or similar too, the opposition between high and low targets, i.e., in opposition to L*. Computational dynamical systems modeling approaches that build on this apparent asymmetry also may provide a valuable extension of the present results; this is work that we are currently pursing (Iskarous, Steffman & Cole, 2023).

In summary, the results presented here show that some distinctions among nuclear tunes are graded and noisy, in line with the idea that they may be modeled as falling along a continuum, e.g., an alignment continuum as discussed in Gussenhoven (1984). This may suggest a model of intonational meaning that allows phonetic variation to play a role. Ladd (2022) in his recent review raises this as a central challenge in the study of intonation: "[. . .] phonological categories of intonation, whatever they are, are subject to meaningful gradient variation [. . .] The fact that an intonationally conveyed pragmatic distinction seems categorical does not entail that the corresponding intonational distinction involves categorically distinct phonological elements. Until we understand this better, our phonological analyses are likely to make spurious categorical distinctions" (p 252). We do not claim to have the answers to these thorny questions of intonational meaning, but our results clearly speak to the view that (1) many claimed distinctions in the AM model do not exhibit robust category-level distinctiveness in F0 form, and (2) this does not preclude meaning-based differences as a function of continuous phonetic variation. We believe that the present line of research into intonational form will be fruitfully complemented by meaning-focused studies that are cognizant of phonetic variation (in the vein of e.g., Calhoun, 2012).

Taking stock of the results and our interpretations outlined above, the present study raises several key questions for future modeling work. Most fundamental is the question of the extent to which discrete, atomic, categories are the best model for intonational forms. This study presents no definitive answer to this deep question, but does show that certain form distinctions are more category-like than others. This suggests, at a minimum, that discreteness is not a given. Two recent approaches to modeling intonational form strike us as appropriate to pursue this observation further. First is the dynamical systems modeling work in Roessig et al. (2019) noted above. In that study, a single acoustic parameter, tonal onglide for pitch accents, was modeled. The authors present a dynamical system with a two-well potential energy function. The wells define an attractor landscape with two attractors, one for a falling pitch accent (negative onglide), and one for a rising pitch accent (positive onglide). The "tilt" of the attractor landscape can be modified by a control parameter which effectively modulates two factors: first, the likelihood that a falling or rising accent will be produced, and second, the steepness or slope of the onglide within those two classes. This can be considered a model of intonational *planning* and *selection* in the sense that it defines a range and likelihood of values for the modeled parameter. The likelihood of the production of a given onglide value is thus determined in the model by the control parameter which could be context- and/or speaker-specific. An obvious appeal of this approach as it relates to our study is that it naturally models both a dichotomous, category-like distinction (via a two-well potential energy function), and continuous variation within and between the rising/falling dichotomy (via the modification of that function by the control parameter). This can be pursued further in its application to other pitch accent distinctions and other measures, or, to whole tunes, though in the latter case, modeling a single parameter representing the whole tune may be reductive.

We believe the aforementioned work of Iskarous et al. (2023) may help serve as a motor for more modeling and theoretical advancement. The authors in that paper develop a dynamical systems model which is a generative model for F0 trajectories, defined by a differential equation with control parameters that define the F0 values, and shape, a particular pitch accent will take. They use the model to describe variation in American English pitch accents, showing that timing and scaling properties in rising pitch accents are well captured, and that, crucially, a dichotomous division between "high" (H*, LH*, L*H) and "low" (L*) accents is emergent through continuous variation in the parameters. This model, unlike that of Roessig et al. (2019) is not one of selection, but rather the *execution* of an F0 movement. Though both of these modeling approaches are relatively new, we think they offer promise, especially if further developed to unite selection and execution, for example the union of a selection model for parameter inputs which are passed to a generative execution model for F0, encoding the idea that selection and execution (here of intonation) are organized by similar computational structures (Iskarous and Pouplier, 2022). Regardless, what both of these approaches readily capture are what we see as two crucial empirical insights. First is that there is variation in intonational form that is fine grained and continuous. Second, is that there is variation that is category-like and discontinuous, leading us to the metaphor of a hierarchy. In this sense, we can say that speakers' production and perception of intonational forms reflect categories (categorically distinct forms) *and* continua. We believe models of intonational representation, production and perception can be fruitfully evaluated in terms of their ability to capture both of these empirical phenomena in a single system.

4.2. Some limitations and future directions

The present study raises many questions for future research. First and perhaps most fundamental is the question of whether and how the presence of pragmatic/discourse-based context may license tune distinctions in production or perception. In presenting tunes in the absence of a meaningful context, the present study leaves unanswered the question of how discourse context may impact the production and perception of a given tune. This is a definite limitation of the present study, which will benefit from being tested in the future. We suggest that future work can attempt to pair these, or other stimuli with written (or potentially, spoken) contexts which are predicted to fit with that tune. Though quantifying the extent to which context should support a given tunes strikes us as quite difficult, this first basic step of testing the presence/absence of a beneficial context would be valuable. If for example, the presence of context enhances particular distinctions, this could constitute evidence that certain tunes "need" context to be well-distinguished, while others, like the well-separated tunes in the present study, do not. If this method is validated it could also be used as a confirmatory test for hypothesized tune functions: if a particular imitative tune distinction is enhanced in the presence of particular discourse contexts, this may be taken to support the hypothesized tune meaning. Future work in this direction has the potential to shed light on these rich and complex questions of tune form and function.

A similar limitation should be considered in light of the fact that particular words or sentences may facilitate the production of particular tunes, as for example the "calling contour", which is natural to produce over names, but less so for other words (i.e., when not calling a person by name). These word-specific and sentence-specific effects on particular tunes remain to be tested and constitute another sort of context that may impact intonational distinctions. We expect that certain contexts, lexical items, or sentences, may enhance or facilitate the production and perception of certain tunes, and yet discovering such effects remains a challenge, as already noted, due to our current incomplete understanding of the pragmatic functions of each tune. An important future direction will be to test the distinctions that emerge when tunes are situated in discourse contexts. This undertaking should go hand in hand with the development of a theory of intonational meaning that applies across each of the predicted AM model tunes and delimits the discourse contexts where each should appear.

Several other methodological considerations provide avenues for future research. One core assumption is that the use of resynthesized F0 as stimuli for the imitation of tunes is adequate to elicit the tune distinctions proposed in the AM model, and specifically, that the resynthesized stimuli allow participants to access their stored cognitive representations of tunes based on their prior experience with the language. Using resynthesized F0 allowed us to develop a set of stimuli with F0 patterns that were as distinct from each other as possible, while still observing the implementation guidelines in Pierrehumbert (1980, 1981) and Veilleux et al. (2006). Making maximally distinct (resynthesized) tunes is beneficial for testing if and how each of these distinctions will be reproduced by speakers. However, if some resynthesized tunes are better representations of the intended category than others, this could result in more tenuous distinctions for poorly represented tunes. As outlined in Section 2.1, our resynthesis was informed by both schematized F0 trajectories of tunes and F0 tracings of natural productions presented in the ToBI training guidelines and in Pierrehumbert (1980). These reference materials lack precise phonological or phonetic landmarks for the alignment of all F0 targets, and in the absence of community-endorsed guidelines, we drew on our experience with ToBI labeling of American English to specify landmarks as anchors for each tone, which were then systematically used for all tunes and all model sentences. Nonetheless, and despite our best efforts, we acknowledge that some listeners may judge the resynthesized stimuli to have been more successful for some tunes than for others. A potentially superior resynthesis method would have been based on clear productions of the 12 tunes by an individual speaker, or by many speakers of the same dialect, but to date no such dataset exists. In our ongoing research extending this line of work we are examining how imitation of natural productions compare to those of resynthesized F0; results of this investigation will help in evaluating the appropriateness of resynthesized tunes for imitative tasks in general.

A further methodological consideration is the reliance on F0 alone to convey tune distinctions, as noted in Section 2.4. Focusing on F0 parameters is certainly standard in intonational research, and the AM model we tested is one that generates distinct F0 trajectories for each tested tune, on the basis of their tonal specification. Nevertheless, it is possible that in manipulating only F0 in the stimuli we have omitted other cues that speakers and listeners use when implementing intonational distinctions such as duration, amplitude, voice quality and articulatory strength. Future work in this line of research will help develop a more holistic understanding of how intonational distinctions are implemented in speech above and beyond F0. Another consideration in the present study is the fact that all tunes were produced over tri-syllabic, stress-initial words. Certain tunes, especially more dynamic tunes such as L*HLH, might be subject to tonal crowding, compression or truncation in this context (e.g., Grabe, Post, Nolan & Farrar, 2000). Though three syllables was, by our assessment, enough material for the distinctions to be heard and produced, it remains an empirical question if more material (i.e., more syllables) would allow additional finer-grained distinctions among tunes to emerge. Cole & Steffman (2023) is another study we have carried out which offers a preliminary answer to this question, though it made use of a different set of tunes (with H* and L* pitch accents) and cannot be directly compared to these results. In that study, listeners heard the stimuli over a tri-syllabic stress-initial word (as in the present study), but then were prompted to re-produce the tune over a word with one, two, three, or four syllables. Cluster analyses on that data found no more emergent clusters for tunes produced over four syllables versus three syllables, and parameters like ending F0 value did not become more distinct across tunes from three to four syllables either, suggesting that having more material over which to produce those tunes did not result in better differentiation. We are currently in the process of extending that study to look at a subset of the tunes tested here, which will speak directly to the question of how varying syllable counts impact the realization of these tunes.

Finally, another important consideration for future work is the potential importance of prenuclear material. Nuclear tunes are "special" in the AM model only in the sense that every intonational phrase will have a nuclear tune, and in the observation that pre-nuclear material seems to be less important for conveying intonational meaning and is sometimes described as "ornamental" (Büring, 2007, though see Bishop, 2017; Braun & Biezma, 2019). However, nuclear tunes cannot be understood fully when analyzed in isolation (e.g., in the absence of prenuclear pitch accents and phrase accents). In the same way that certain nuclear tune configurations may be more likely, certain relationships between pre-nuclear and nuclear material are likely more common, and thus potentially more accessible to speakers. One well-known case of this would be the so-called "hat pattern" composed of a series of high pitch accents. Our study, lacking pre-nuclear accents, cannot address this question of the importance of prenuclear material, or the relationship between pre-nuclear and nuclear material for speakers' representation and production of tunes. Future work should thus consider how pre-nuclear patterns are produced in this experimental paradigm and consider possible facilitatory (or inhibitory) influences of pre-nuclear tones on following nuclear tunes.

4.3. Conclusions

In summary, the present results provide an exploration of the nature of distinctions among 12 American English nuclear tunes. As outlined in Section 1, the proposed distinctions from the AM model have not previously been investigated in a systematic fashion. Prior research has often focused on the distinction between a single pair of tunes, and usually in a single edge tone context (or, not controlling for edge tone context). Our aim was to provide a rigorous test for the distinctions in intonational form predicted by the AM model, in a more systematic fashion: across controlled edge tone contexts and comparing among a large set of tunes. Our results suggest that all the distinctions predicted by the AM model are present to some degree, but that certain distinctions are primary and exhibit category-like behavior, while others are smaller and more variable, and may be understood as representing structured phonetic variability within categories, conditioned in part by intonational context. In other words, the results reported here are consistent with a system of distinctions in intonational form that is hierarchical and context dependent.

CRediT authorship contribution statement

Jeremy Steffman: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Jennifer Cole:** Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing, Investigation. **Stefanie Shattuck-Hufnagel:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Chun Chan for technical support. We are additionally grateful for valuable feedback on this work from the reviewers, members of the Prosody & Speech Dynamics Lab and Phonatics Discussion Group at Northwestern University, and from attendees at *Speech Prosody 2022.* This project was supported by NSF BCS-1944773 (Cole, PI).

Appendix

Appendix 1:. More details for the TCoG computations

The equation in (2) below is used to compute TCoG in a given interval, in this case, in the window of the nuclear word.

$$\textit{Temporal TCoG} = \frac{\sum_{i=1}^{n} F0_i t_i}{\sum_{i=1}^{n} F0_i} \quad (2)$$

As shown in (2), temporal TCoG (henceforth just "TCoG") is derived by multiplying a given time measure by the F0 value at that time, and summing these weighted values across an interval from t_i to t_n, then dividing by F0 summed across the same interval. The key insight is that regions with higher F0 contribute more to the weighted average, essentially drawing the temporal TCoG measure towards them (in time).

To illustrate how this computation works, consider the diagram in Fig. A1 (similar to examples in Barnes et al., 2021). Fig. 3, Panel A shows a schematized F0 rising-falling movement which extends over 400 ms, and ranges from 50 to 150 Hz. Note that the peak location always occurs at precisely 200 ms. Panel A shows three ways the peak can be approached: a linear rise, a rise that begins with a steeper slope, which we will refer to as "domed" and a rise that begins with a flatter slope before rising more rapidly to the F0 peak, which we will refer to as "scooped". The Temporal TCoG for these three rises was computed using the formula in (2) and is plotted by the dashed vertical lines in the panel. Note that for the domed rise, the higher F0 achieved earlier in the window effectively pulls temporal TCoG earlier in time, as those higher and earlier values contribute more to the weighted average. The opposite is true for the scooped rise. Panel B shows an analogous effect of the shape of the fall from an F0 peak: a domed fall pulls TCoG later in time while a scooped fall pushes it earlier. Fig. A1, panel C shows a monotonically rising F0 pattern where the same generalization holds in terms of the influence of rise shape: a scooped rise pushes TCoG later, while a domed rise pulls it earlier in the interval.

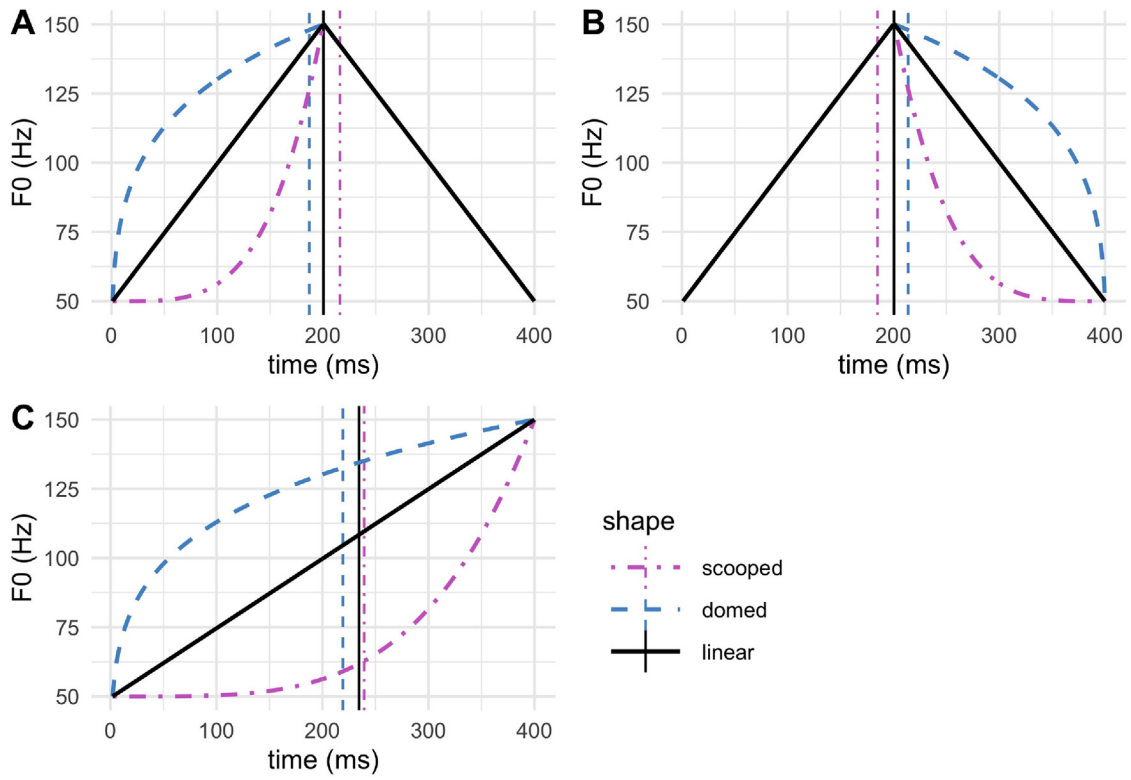

Fig. A1. Examples of how rise shape and fall shape influence the location of temporal TCoG (see text).

Appendix 2:. Additional figures

(See Fig. A2, Fig. A3, Fig. A4)

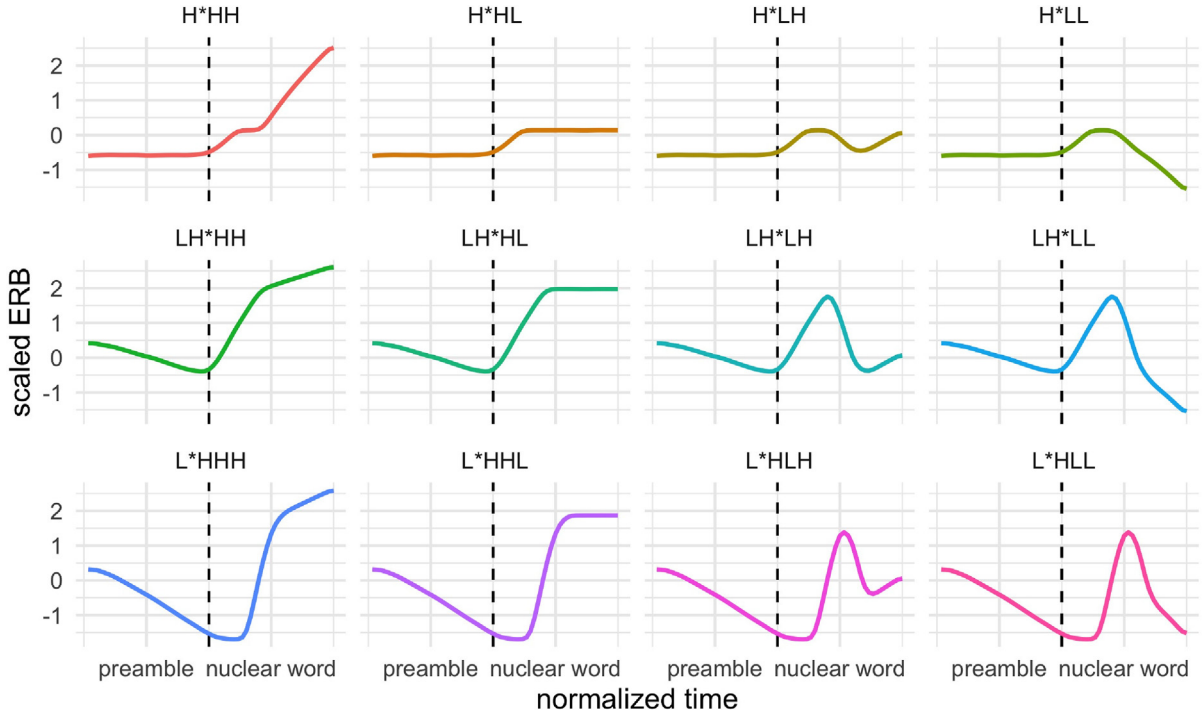

Fig. A2. Model stimuli in scaled F0, averaged across the four models for each tune. The vertical line separates the preamble from the nuclear word. Note that the preamble varies only across pitch accents (rows).

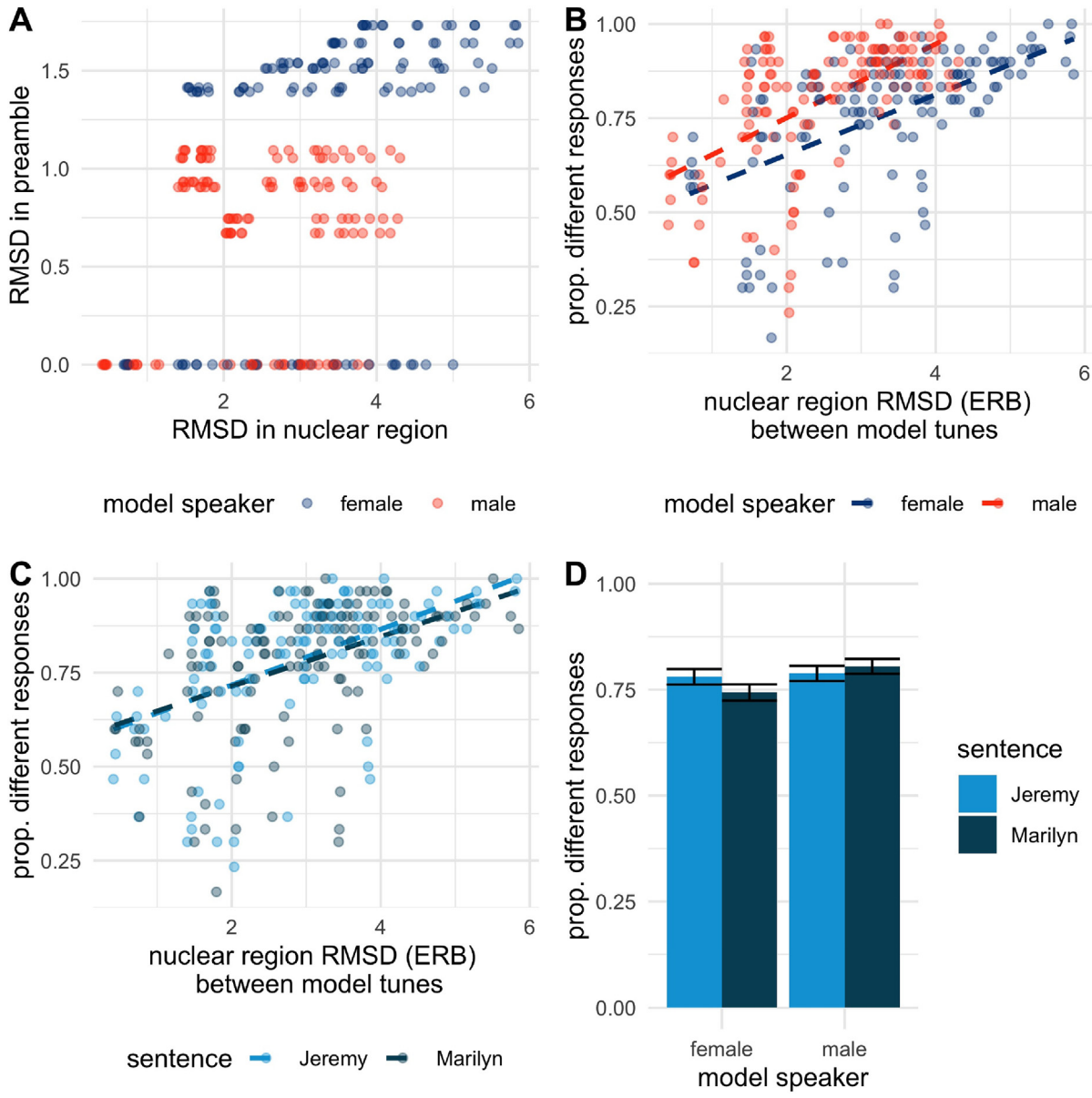

Fig. A3. Plots showing aspects of the speech perception data. Panel A shows the relation between nuclear and preamble RMSD in the stimuli, also colored by model speaker. Note that variation in darkness in points indicates that multiple points are stacked on top of one another. The remaining plots showing discrimination accuracy as a function of model speaker (panel B), model sentence (panel C), both plotted against nuclear region RMSD on the x axis. Panel D shows discrimination accuracy which is collapsed by RMSD and split by sentence (bar color) and model speaker (x axis).

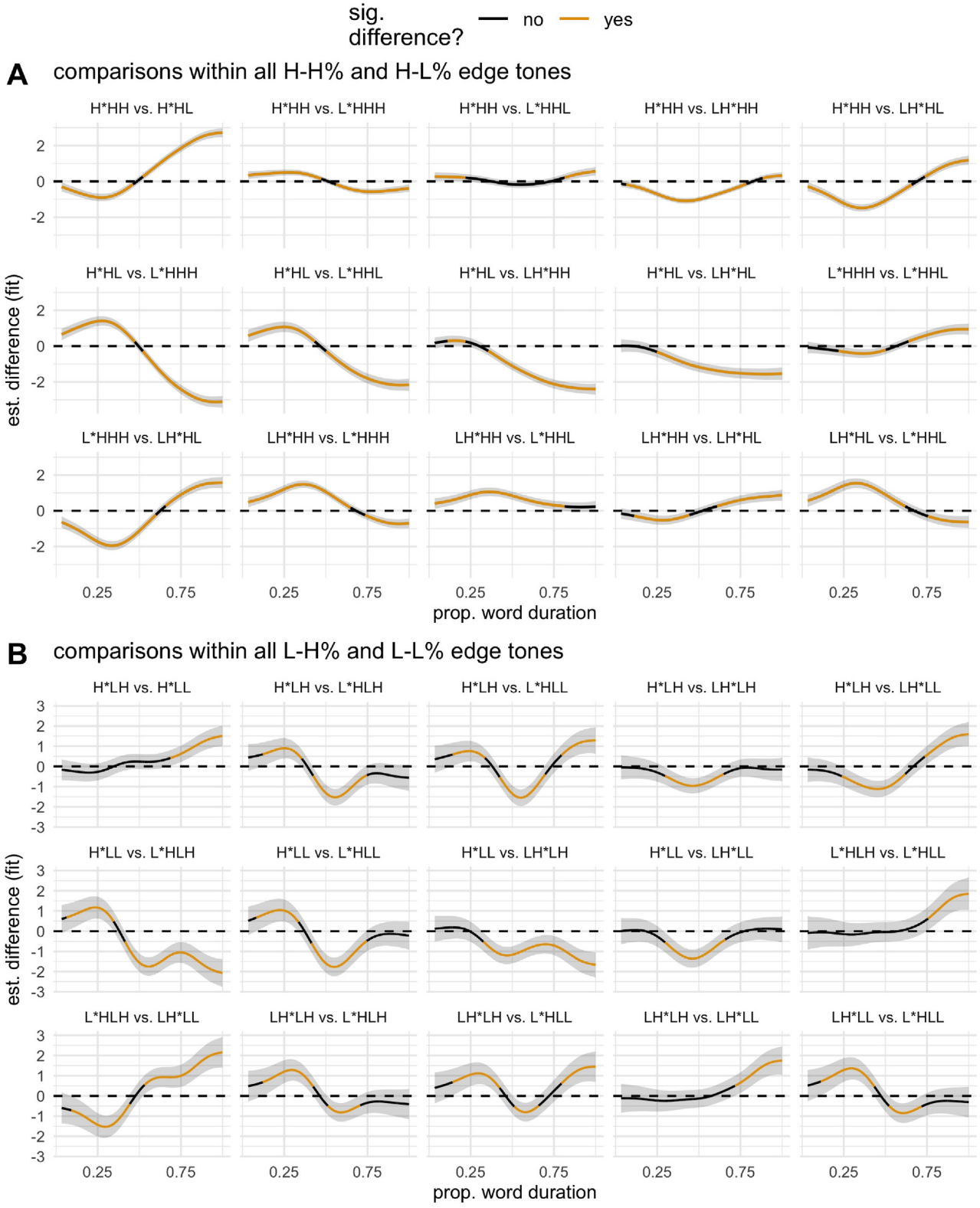

Fig. A4. Difference smooths for pairwise comparisons of tunes. Panel A shows all comparisons between tunes with H-H% and H-L% edge tones (panels A and B in Fig. 4). Panel B shows all comparisons between tunes with L-H% and L-L% edge tones (panels C and D in Fig. 4). Color indicates if a difference is significant at a given point in normalized times.

References

Arvaniti, A., & Garding, G. (2007). Dialectal variation in the rising accents of American English. *Papers in Laboratory Phonology, 9*, 547–576.

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology, 3*(2), 337–383.

Barnes, J., Brugos, A., Veilleux, N., & Shattuck-Hufnagel, S. (2021). On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics, 85* 101020.

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation, 3*(30), 255–309.

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

Bishop, J. (2017). Focus projection and prenuclear accents: Evidence from lexical processing. *Language, Cognition and Neuroscience, 32*(2), 236–253.

Braun, B., Kochanski, G., Grabe, E., & Rosner, B. S. (2006). Evidence for attractors in English intonation. *The Journal of the Acoustical Society of America, 119*(6), 4006–4015.

Braun, B., & Biezma, M. (2019). Prenuclear L* + H activates alternatives for the accented word. *Frontiers in psychology, 10*, 1993.

Büring, D. (1997). *The meaning of topic and focus—The 59th Street Bridge accent*. London: Routledge.

Büring, D. (2007). Intonation, Semantics and Information Structure. In G. Ramchand & C. Reiss (Eds.), *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press.

Bürkner, P. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal, 10*(1), 395–411.

Burdin, R. S., & Tyler, J. (2018). Rises inform, and plateaus remind: Exploring the epistemic meanings of "list intonation" in American English. *Journal of Pragmatics, 136*, 97–114.

Burdin, R. S., Holliday, N. R., & Reed, P. E. (2022). American English pitch accents in variation: Pushing the boundaries of mainstream American English-ToBI conventions. *Journal of Phonetics, 94* 101163.

Calhoun, S. (2004). Phonetic Dimensions of Intonational Categories-the case of L+ H* and H.

Calhoun, S. (2012). The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics, 40*(2), 329–349.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods, 3*(1), 1–27.

Chodroff, E., & Cole, J. (2019a). In *Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English* (pp. 1966–1970). International Speech Communication Association.

Cole, J., & Steffman, J. (2023). Enhancement of intonational contrasts in American English. *Proceedings of the International Congress of Phonetic Sciences, 2023*, 1370–1374.

Chodroff, E. & Cole, J. (2019b). The phonological and phonetic encoding of information status in American English nuclear accents. In *Proceedings of the 19th International Congress of Phonetic Sciences*. York.

Cole, J. & Shattuck-Hufnagel, S. (2011). The phonology and phonetics of perceived prosody: What do listeners imitate? *Proceedings of INTERSPEECH 2011,* 969-972. International Speech Communication Association.

Cole, J., Steffman, J., Shattuck-Hufnagel, S., & Tilsen, S. (2023). Hierarchical distinctions in the production and perception of nuclear tunes in American English. *Laboratory Phonology, 14*(1).

Dainora, A. (2001). *An empirically based probabilistic model of intonation in American English*. University of Chicago. Doctoral dissertation.

Dainora, A. (2006). Modeling intonation in English: A probabilistic approach to phonological competence. *Laboratory Phonology, 8*, 107–133.

Dilley, L. C. (2005). *The phonetics and phonology of tonal systems (Doctoral dissertation*. Massachusetts Institute of Technology).

Dilley, L. C. (2010). Pitch range variation in English tonal contrasts: Continuous or categorical? *Phonetica, 67*(1–2), 63–81.

Dilley, L. C., & Heffner, C. C. (2013). The role of F0 alignment in distinguishing intonation categories: Evidence from American English. *Journal of Speech Sciences, 3*(1), 3–67.

D'Imperio, M. (2000). *The role of perception in defining tonal targets and their alignment*. The Ohio State University.

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software, 65*(4), 1–34.

Gerrits, E., & Schouten, M. E. (2004). Categorical perception depends on the discrimination task. *Perception & psychophysics, 66*, 363–376.

Grabe, E., Post, B., Nolan, F., & Farrar, K. (2000). Pitch accent realization in four varieties of British English. *Journal of Phonetics, 28*(2), 161–185.

't Hart, Johan. (1991). F0 stylization in speech: Straight lines versus parabolas. *Journal of the Acoustical Society of America, 90*(6), 3368–3370.

Hermes, D. J. (1998). Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research, 41*(1), 73–82. https://doi.org/10.1044/jslhr.4101.73.

Im, S., Cole, J., & Baumann, S. (2023). Standing out in context: Prominence in the production and perception of public speech. *Laboratory Phonology, 14*(1).

Iskarous, K., Steffman, J., & Cole, J. (2023). American English Pitch Accent Dynamics: A Minimal Model. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1469–1473).

Iskarous, K., & Pouplier, M. (2022). Advancements of phonetics in the 21st century: A critical appraisal of time and space in Articulatory Phonology. *Journal of Phonetics, 2* (95) 101195.

Jeong, S. (2018). Intonation and sentence type conventions: Two types of rising declaratives. *Journal of Semantics, 35*(2), 305–356.

Jun, S. A. (2005). *Prosodic typology: The phonology of intonation and phrasing*. OUP Oxford.

Jun, S. A. (2014). *Prosodic typology II: The phonology of intonation and phrasing*. OUP Oxford.

Kaland, C. (2021). Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *Journal of the International Phonetic Association*.

Kawahara, H., Cheveigné, A. D., Banno, H., Takahashi, T., & Irino, T. (2005). Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. *In Ninth European Conference on Speech Communication and Technology*.

Knight, R. A. (2008). The shape of nuclear falls and their effect on the perception of pitch and prominence: Peaks vs. plateaux. *Language and Speech, 51*(3), 223–244.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Ladd, D. R. (2022). The trouble with ToBI. In Barnes, J. & Shattuck-Hufnagel (Eds.). *Prosodic theory and practice*, 247-258.

Ladd, D. R., Mennen, I., & Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *The Journal of the Acoustical Society of America, 107*(5), 2685–2696.

Ladd, D. R., & Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of phonetics, 31*(1), 81–112.

Lenth, R. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. *R package version*, 1.7.1-1 https://CRAN.R-project.org/package=emmeans.

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software, 4*(40), 1541. https://doi.org/10.21105/joss.01541.

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2020). *Estimation of Model-Based Predictions*. CRAN: Contrasts and Means.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech, 2017*, 498–502.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*(5–6), 453–467.

Niebuhr, O., d'Imperio, M., Fivela, B. G., & Cangemi, F. (2011). Are There "Shapers" and "Aligners"? Individual Differences in Signalling Pitch Accent Category. *In ICPhS*, 120–123.

Pierrehumbert, J. B., & Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica, 46*(4), 181–196.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation (Doctoral dissertation*. Massachusetts Institute of Technology).

Pierrehumbert, J. (1981). Synthesizing intonation. *The Journal of the Acoustical Society of America, 70*(4), 985–995.

Remijsen, B., & van Heuven, V. J. (1999). Gradient and categorical pitch dimensions in Dutch: diagnostic test. In *Proceedings of the 14th International Congress of Phonetic Sciences* (Vol. 2, pp. 1865-1868).

Roessig, S., Mücke, D., & Grice, M. (2019). The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PLoS One1, 14*(5), e0216859.

Rosenberg, A. (2010). AuToBI-a tool for automatic ToBI annotation. *In Eleventh Annual Conference of the International Speech Communication Association*.

Rudin, D. (2022). Intonational commitments. *Journal of Semantics, 39*(2), 339–383.

Schneider, K., & Lintfert, B. (2003). Categorical perception of boundary tones in German. In *In Proceedings of the 15th International Conference of the Phonetic Sciences* (pp. 631–634).

Schweitzer, A., & Möbius, B. (2009). Experiments on automatic prosodic labeling. *In Tenth Annual Conference of the International Speech Communication Association*.

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. *In Proceedings of ICPhS XVII*, 1846–1849.

Silverman, K. E., & Pierrehumbert, J. B. (1987). The timing of prenuclear high accents in English. *The Journal of the Acoustical Society of America, 82*(S1), S19–S.

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics, 84*, 101017.

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

Steffman, J., & Cole, J. (2022). An automated method for detecting F0 measurement jumps based on sample-to-sample differences. *JASA Express Letters, 2*(11) 115201.

Sundberg, J. (1973). Data on maximum speed of pitch changes. *Speech Transmission Lab. Quarterly Progress Status Report., 4*, 39–47.

Syrdal, A., McGory, J., 2000. Inter-transcriber reliability of ToBI prosodic labeling. Paper presented at the International Conference on Spoken Language Processing (ICSLP), Beijing, China.

Tilsen, S., Burgess, D., & Lantz, E. (2013). Imitation of intonational gestures: A preliminary report. *Cornell Work. Pap. Phon. Phonol*, 1–17.

van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2016). Itsadug: Interpreting time series and autocorrelated data using GAMMs [R package].

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of phonetics, 71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413–1432.

Vehtari A., Gabry J., Magnusson M, Yao Y., Bürkner P., Paananen T., Gelman A .(2020). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.(2016). *R package version 2.4.1.*

Veilleux, N., Shattuck-Hufnagel S. & Brugos A. 6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI. January IAP 2006. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Wightman, C., & Campbell, N. (1995). *Improved labeling of prosodic structure*. Speech Audio Process: IEEE Trans.

Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on speech and audio processing, 2*(4), 469–481.

Wightman, C., Syrdal, A., Stemmer, G., Conkie, A., & Beutnagel, M. (2000). Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative speech synthesis. *In: Proc. ICSLP., 2*, 7174.

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+ H*. *Cognitive science, 32*(7), 1232–1244.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC.