Steffman, Jeremy & Cole, Jennifer. 2024. Metrical enhancement in American English nuclear tunes. *Glossa: a journal of general linguistics* 9(1). pp. 1–44. DOI: https://doi.org/10.16995/glossa.15297

Open Library of Humanities

Metrical enhancement in American English nuclear tunes

Jeremy Steffman, The University of Edinburgh, UK, jeremy.steffman@ed.ac.uk

Jennifer Cole, Northwestern University, US, jennifer.cole1@northwestern.edu

We present two experiments aimed at testing the nature of intonational categories through the lens of enhancement. In an imitative speech production paradigm, speakers heard a model intonational tune and were prompted to reproduce that tune on a new sentence in which the syllable count of the word carrying the tune varied. Using the prevalent auto-segmental metrical model of American English as a basis for potential tune categories, we test how distinctions among tunes are enhanced across different metrical structures. First, with a clustering analysis, we find that not all predicted distinctions are emergent. Secondly, only the largest distinctions, those that emerge in the clustering analysis, are enhanced as a function of metrical structure. Measurable differences between tunes which cluster together are detectable, but critically, are not enhanced. We discuss what these results mean for the nature and number of intonational categories in the system.

1 Introduction

The study of intonational phonology has long been concerned with the question of intonational categories: what are the phonological atoms (for a given language), and how are they manifested in the speech signal? This question is a necessary complement to the study of the function of intonation as a phonological system and its role in conveying linguistic meaning in speech. However, even in the relatively well-studied intonational system of American English, there remain longstanding and continued questions about the number and fundamental nature of intonational categories (e.g., Gussenhoven 1984; Dilley & Heffner 2013; Ladd 2022). In the present study we pursue a fresh approach to the question of intonational categories in American English. In analogy to phenomena in the segmental domain, we examine intonational categories through the lens of enhancement. With intonation elicited using an imitative speech production paradigm, in which participants hear and reproduce an intonational melody, we examine intonational melodies realized across various metrical structures (primarily varying syllable count) for evidence that predicted phonological distinctions are, or are not, enhanced. In what follows we motivate the general research question, define and operationalize enhancement, and describe the logic of the experiments presented here.

1.1 Intonational Phonology and Intonational Categories

One common framework for understanding and modeling intonational phonology is Auto-segmental Metrical (henceforth AM) Theory. The phonological atoms in the theory are high (H) and low (L) tones which, depending on the language, serve different functions in marking prominence and boundaries in speech. In the prevalent AM model of Mainstream American English (MAE) (Pierrehumbert 1980; Beckman & Pierrehumbert 1986), H and L tones may associate with lexically stressed syllables in words with phrase-level prominence. These intonational features are called “pitch accents”, and are indicated as H* and L*. Pitch accents can also have two component tones, in which case a single tone is primarily associated with the pitch-accented syllable, and the other tone immediately precedes or follows it. For example, in the bitonal L+H* pitch accent, a low target precedes the H* accent on the stressed syllable. In the bitonal L*+H accent, there is a low (L*) target in the stressed syllable, which is immediately followed by rise (H) after the stressed syllable.

The object of the present study is the *nuclear tune*, which consists of the final or “nuclear” pitch accent in an intonational phrase, followed by two edge tones that function to mark the end of prosodic phrasal domains. The AM model of MAE recognizes two layers of prosodic phrase structure. The higher level is the intonational phrase, which comprises one or more lower-level “intermediate” phrases. As such, the end of every intonational phrase is also the end of the rightmost (and possibly the only) intermediate phrase it contains. The first edge tone in the nuclear tune is the “phrase accent”, labeled as H- or L-, which marks the right edge of the intermediate phrase and defines the pitch target for the interval between the nuclear pitch accent and the end of the phrase. This interval may comprise one or more syllables, depending on the location of the nuclear pitch accent relative to the end of the intermediate phrase. The second edge tone in the nuclear tune is the "boundary tone", labeled H% or L%, marking the end of the intonational phrase and situated on the phrase-final syllable. A nuclear pitch accent is obligatory within each intermediate phrase, as are the phrase accent and boundary tone at the end of the intermediate and intonational phrases. Thus, this system generates a nuclear tune with these three components, e.g., H*L-L%, in every intonational phrase, as the minimal intonational melody of speech. An intonational phrase may include other features as well. For example, there may be additional "pre-nuclear" pitch accents marking phrasal prominence on words preceding the nuclear accented word, and an additional phrase accent for each intermediate phrase preceding the rightmost one in the intonational phrase.

The AM model of American English comprises an inventory of five pitch accents (setting aside the downstepped High tone: !H), two phrase accents, and two boundary tones, each of which may freely combine with one another to create an inventory of 20 nuclear tunes. These tunes are considered phonologically distinct on the basis of differences in their tonal specification. The model is thus very explicit about the phonological inventory of the language and possible tone sequences. However, there exists little empirical work testing the model's inventory of tones and tone sequences. Accounts of phonological distinctiveness making reference to meaning distinctions exist for some but not all tunes in the model's set (e.g., Pierrehumbert & Hirschberg 1990). However recent work has shown a non-deterministic and many-to-many mapping between intonational categories and pragmatic and discourse functions (e.g., Chodroff & Cole 2018; 2019; Im et al. 2023). One additional, and fundamental, challenge to the categorically distinct intonational atoms proposed in the AM model, is the large range of variation in their realization in F0. A common approach to studying this has been to examine the difference between just two tones or tunes, e.g., between H* and L+H*. Though the model claims these as two categorically distinct entities, expert annotators struggle to reliably differentiate them, and the acoustic parameters that distinguish them have been questioned (Ladd & Schepman 2003; Dilley 2005; Calhoun 2012). An alternative proposal is that they represent variants of one single intonational category (Ladd 2022). Similar difficulties have emerged in the study of tonal alignment, where the model posits L+H* and L*+H as two categorically different entities, with an alternative proposal by Gussenhoven (1984) that this distinction represents variants along an alignment continuum.

Underpinning both of these controversial cases is what Ladd (2022) identifies as "gradience" (following Bolinger 1961). Unlike the categorical difference in lexical meaning conveyed by cues to segmental contrasts, in MAE the meaning distinctions conveyed by intonation can be seen in some if not all cases as falling along a gradient, scalar dimension, e.g., "emphasis".

Even in the case where the pragmatic function conveyed by intonation seems categorical, Ladd notes that this "[…] 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). In this view L + H* and H* could be described as belonging to a single phonological category: "high accent". Variation in the timing and scaling of the rise in this accent may indeed signal various pragmatic nuances, a case of phonetically meaningful within-category variation. Our present study does not address intonational meaning, and so will not provide a full assessment of this alternative understanding of intonational categories and meaning. However, the notion that certain distinctions in intonational form may entail categorical differences, while others may entail (meaningful) gradient variation presents a view of intonational phonology that underpins much of the controversy surrounding the AM model for MAE.

A more general challenge is simply that there has not been, to our knowledge, a study that shows that each of the proposed categories is measurably distinct from other categories, in phonetic form. Many studies focus on just two pitch accents or just two tunes/tune shapes (e.g., rising, versus falling F0). Somewhat strikingly, several recent studies that have examined the intonational system more holistically (testing eight tunes, or twelve tunes), do *not* find compelling evidence that all of the proposed distinctions are robust and well-maintained by speakers and listeners (Steffman et al. 2022; 2024; Cole et al. 2023). Here we further this line of inquiry testing the predicted distinctions from the MAE AM model. In summary, our reading of the literature leads us to conclude that the inventory of tones and the tunes that they compose is not a settled issue, with some evidence against the full proposed inventory. With this we will now turn to the notion of enhancement and its possible use in examining the nature of intonational categories.

1.2 Phonological categories and enhancement

We use the term enhancement in this paper as follows: enhancement occurs when some "intrinsic" property of a phonological category becomes more prominent in the speech signal, or when the distinction between two categories becomes larger, this latter effect being contrast enhancement. Note that these two notions can be related, e.g., enhancement of intrinsic properties can entail contrast enhancement.

This topic has long been considered in the segmental literature. One line of research in this vein has been through the lens of *enhancement theory* (Stevens & Keyser 1989; Keyser & Stevens 2006), couched in distinctive feature theory, which describes enhancement as targeting the distinctive features that differentiate phonological categories. For example, in English /ʃ/ is generally produced with lip rounding, while /s/ is not. Because rounding is not otherwise contrastive in fricatives in English, the rounding distinction between these two fricatives makes sense when we consider that it enhances the [-anterior] feature of /ʃ/, which marks its contrast with the [+anterior] /s/ (see Keyser & Stevens 2006 for various other examples). The core idea of enhancement theory is thus that the properties that make phonological categories distinct are targeted for enhancement, in this case by additional or "secondary" articulations that produce a feature-defining acoustic output. One other thread of research in the phonetics literature examines enhancement under prosodic prominence. The central finding in this literature is that in positions of prosodic prominence (e.g., the lexically stressed syllable of a word with phrasal stress, associated with a pitch accent), cues that differentiate phonological categories are enhanced. (Note that this type of enhancement could also be analyzed in terms of distinctive features.)

The phonetics literature has documented data commensurate with enhancement in speech acoustics in a variety of domains (e.g., Cho 2004; 2005; De Jong 2004; Cole et al. 2007; Beckman et al. 2011; Garellek 2014; Cho et al. 2017; Kim et al. 2018). For example, voice onset time (VOT) for word-initial voicing contrasts in American English becomes more distinct across voicing categories when prosodically prominent (Cole et al. 2007; Kim et al. 2018), and VOT differences for the voicing contrast in Central Standard Swedish become more pronounced at slower speaking rates (Beckman et al. 2011). Similar enhancement effects conditioned by prominence are seen in spectral measures of vowels, which in prominent positions are co-articulated less with adjacent vowels, reducing the variation in vowel formants from the coarticulation-inducing vowel. For instance, under prominence, the high front vowel /i/ is produced with a more peripheral articulation resulting in greater distance from nearby vowels in F1 × F2 formant space (Cho 2005). As this example illustrates, enhancement under prosodic prominence serves to enhance phonological distinctions by targeting properties that differentiate phonological categories, such as the height and backness of a vowel, or by cueing features that are central to a category.

In the realm of intonation research, the idea that F0-based distinctions among intonational features can be modulated as a function of phonological context is not new. In comparison to work on segmental enhancement, much of this research frames the question around a loss or obfuscation of F0-based distinctions, particularly under time pressure, i.e., when the duration of segmental carriers is shorter, or there is less segmental material to carry a tune. We will refer to these as reducing contexts. *Truncation* is one process observed in reducing contexts, where the ending part of an F0 contour is "cut off" or truncated. Another reductive process is *compression*, where the F0 contour is compressed in time, retaining its shape characteristics without truncation of the ending portion. Truncation and compression effects are documented in the literature and have been shown to be language- and dialect-specific (Grabe 1998; Grabe et al. 2000; Yu & Zahner 2018; Sadeghi 2023). In some cases, these effects are described as neutralizing certain intonational distinctions (e.g., Grice 2017), which as shown by Rathcke (2013) for truncation, can lead to difficulty for listeners' identification of intonationally distinct forms, even if neutralization is not complete. While work to date on intonational reduction is limited in scope, there is much less attention paid to intonation production in enhancing contexts. Specifically, we are aware of no study that systematically compares F0 trajectories in enhancing contexts, i.e., those that afford more than the minimal phonological material needed to anchor each tone in each intonation feature, with a context that provides just sufficient phonological material.

1.3 Goal and premise of the present study

To summarize the discussion so far, we have observed that there is a lack of consensus in the literature about the nature and number of phonological distinctions in the intonational phonology of American English. We have observed that the phonetic implementation of intonational features varies as a function of phonological context, in particular, the syllabic and durational properties of the segmental string that anchors intonational features. This variation thus contributes to the challenge in identifying intonational categories in the language. Finally, we have reviewed the view of enhancement as targeting phonetic properties that are important for encoding phonological contrasts, noting that to date there is little work examining enhancement in intonation.

These observations set the stage for the present study. By examining intonation in enhancing contexts we may seek evidence about which F0 properties of an intonation feature or tune serve to define intonational contrasts. Further, by comparing F0 trajectories in reducing contexts with those in enhancing contexts, we may gain insight into whether and how intonational contrasts are neutralized.

Beyond the immediate implications for the MAE intonational model, this study has the potential to inform on intonational phonology more broadly. Unlike lexical-level contrasts, post-lexical intonational systems operate in a meaning/contrast space that is more difficult to define, and can be gradient in a way that lexical contrasts cannot, as described above. Thus one fruitful avenue for better understanding intonational phonology at large, in our view, is the study of enhancement of various intonational elements or tunes composed of them (cf. Arvaniti & Garding 2007). In other words, by examining which distinctions speakers chose to enhance, or don't enhance, we may be able to glean insight into the nature of the intonational system under investigation. From this perspective, we aim to offer a methodological contribution to the study of intonational phonology, and potentially other systems (e.g., lexical tone), via the understanding of which elements are subject to enhancement.

We pursue the question of enhancement through the analysis of F0 trajectories of nuclear tunes in phonologically short contexts where truncation or compression of tunes may occur and in longer contexts where enhancement is possible. Ultimately, we are interested in how the distinctions among tunes are phonetically manifested in reducing and enhancing metrical structures, as a probe for category status. Specifically, if differences in the F0 trajectories for a pair of tunes are enhanced in longer contexts relative to shorter ones, we consider this as evidence for a category-level distinction between the two tunes.

2 Methods

All data, scripts, and stimuli can be found online, hosted on a repository on the OSF at https://osf.io/dbq4w/.

2.1 Experimental methods and tune selection

In the present study, we elicited tunes in a number of metrical contexts using imitative speech production. Imitation has been used frequently to study intonation (e.g., Pierrehumbert & Steele 1989; Braun et al. 2006; Dilley & Heffner 2013; D'Imperio & German 2015; Zahner-Ritter et al. 2022; Cole et al. 2023; see Gussenhoven 2006 for discussion), and offers a tool to elicit the production of a tune with a relatively high degree of control and without reference to tune meaning or function. We use an experimental paradigm of "indirect" imitation of intonational melodies: listeners hear a model utterance that instantiates a tune over a trisyllabic word (described in detail below). They then transpose the tune onto a new word, which contains one, two, three or four syllables. The imitation is "indirect" in the sense that the words in the model utterance and produced utterance are different. The one and two syllable words represent a possible context for truncation or compression as the same tune is realized over fewer syllables. The four syllable words represent an opportunity for enhancement as more material may allow for distinctions between some tunes to be made larger and more apparent than they appear in the trisyllabic stimuli presented to participants. Here we note that enhancement for intonational contrasts can be seen as the other side of the coin of truncation and compression, that is, important cues becoming more salient, distinct, or recoverable, as a function of more material over which they may be realized. Enhancement and reduction are assessed in the imitated productions of tunes through a set of analyses, described below. We first apply a clustering analysis to the unlabeled imitations to determine the number and type of distinct F0 trajectories that emerge from speakers' imitations of a set of phonologically distinct tunes, comparing the number of clusters that emerge in shorter vs. longer metrical conditions. We then proceed with a quantitative analysis of F0 differences among the imitated tunes, looking for F0 differences that are enhanced in longer contexts, and/or reduced in shorter ones. These analyses are described further below.

We tested two sets of six tunes, over two experiments. Our choice of testing only six tunes within an experiment was largely a practical one: given the five metrical structure conditions we tested (described below), and the desire for a full within-subjects manipulation of each factor, we were limited in the number of trials a given participant could reasonably produce.

There are eight nuclear tunes that can be created by combining monotonal H* and L* pitch accents with the two phrase accents and the two boundary tones. The first set of six in tunes in this paper were selected as a subset of these eight tunes, and will be referred to as the tunes in the monotonal accent experiment. Note that in what follows we suppress the edge tone diacritics of – and %, such that, for example, H*H–H% is written as H*HH. The six tunes we selected to test here were informed by previous work (Cole et al. 2023), which identified three pairs of tunes that were poorly distinguished in production (based on data from time-series clustering analysis and neural net classification) and in perception (based on AX perceptual discrimination data). The three pairs of "confusable" tunes were: {H*HH, H*HL}, {H*LH, H*LL}, and {L*HL, L*LH}, where brackets group tunes in a confusable pair together. **Figure 1**, which is described in more detail below, provides schematic representation of each of these tunes. Productions of both {H*HH, H*HL} exhibit a monotonically rising F0 and are expected to be differentiated in the ending F0, with a H boundary tone (the final tone in the sequence) leading to higher F0 at the end of the tune compared to a L boundary tone. These two tunes were shown to be perceived and produced as different from other tunes in Cole et al. (2023), however they were not well distinguished from one another. The case is the same for the pair {H*LH, H*LL}, which also differ in ending F0, and which both contain a H* pitch accent peak with a subsequent fall. Both tunes in the pair {L*HL, L*LH} have a low L* pitch accent realized in an F0 valley, and differ principally in the timing of the rise after the L*. For each of these three pairs, (Cole et al. 2023) found measurable acoustic differences, for example in the timing of the F0 valley for {L*HL, L*LH} or the ending F0 value for {H*HH, H*HL} and {H*LH, H*LL}. However, these differences were small and the tunes in each pair did not readily emerge as distinct in clustering or classification analyses. For this set of tunes, we ask if small differences of this sort are minimized or enhanced in relation to varying metrical structures. If, for example, differences in ending F0 in {H*HH, H*HL} and {H*LH, H*LL} are enhanced, this would suggest that a distinction in ending F0 constitutes an important component of the phonological distinction between the tunes in each pair, which, however, might have been diminished based on the materials used by Cole et al. (2023), where tunes were always realized over a tri-syllabic stress-initial word.

The second set of six tunes comes from another previous study which used the same intonation imitation paradigm, focused on two bitonal pitch accents in the AM model: L+H* and L*+H. This will be referred to as the bitonal accent experiment. These were paired with three edge tones: HH, LH and LL. The selection of these six tunes was also informed by another previous study, Steffman et al. (2024). That study found that, like the monotically rising F0 patterns in Cole et al. (2023), when these rising pitch accents are followed by a H phrase accent and H boundary tone {LH*HH, L*HHH}, the resulting tunes were grouped in the clustering analysis together with other tunes that have rising F0 movements, and were poorly differentiated from one another. Tunes formed with the bitonal pitch accents followed by LH and LL edge tones result in a rising-falling F0 movement, and these four rising-falling tunes were not well distinguished in the clustering analysis presented in Steffman et al. (2024).

Both sets of tunes (with monotonal or bitonal pitch accents) thus test certain distinctions predicted by the AM model that we expect to be large and robust, based on prior work, and other distinctions which are small and less robust.

As noted above, the imitative paradigm involves an “indirect imitation” in which the participants produce a different sentence from those heard as stimuli. The two model stimulus sentences were also presented in different voices on each trial, one sentence in each of two model speaker voices. The model sentences we used were “He answered Jeremy” and “Her name is Marilyn”. Participants then received an orthographic prompt to produce a new sentence, which was one of the ten sentences shown in **Table 1**. These ten sentences represent 5 metrical conditions, which varied the metrical structure of the nuclear (final) word to contain one through four syllables, where each word has initial stress. A further distinction is made between conditions 4a and 4b: the nuclear words in the 4b conditions contain a secondary stress on third syllable, whereas those in 4a have no secondary stress.

Metrical condition	**Sentence**	**Nuclear word metrical structure**
1	He ran with *Moe*	*σ́*
	She lived with *Neil*	*σ́*
2	Her roommate *Nóra*	*σ́σ*
	His neighbor Mánny	*σ́σ*
3	She gathered *lávender*	*σ́σσ*
	They honored *Mélanie*	*σ́σσ*
4a	He went there *críminally*	*σ́σσσ*
	She travelled *mínimally*	*σ́σσσ*
4b	They saw the *nóminàtor*	*σ́σσ̀σ*
	He was a *lúminàry*	*σ́σσ̀σ*

Table 1: The ten target sentences produced by speakers. There are two sentences in each metrical condition, and the same sentences are used in both experiments. The nuclear word is italicized, and the stress pattern and syllable count of the nuclear word is represented at right.

2.2 Stimuli

The same base files were used for generating stimuli in both experiments. There were four of these in total (two speakers producing the two sentences). These were recorded in a soundattenuated booth with a Shure SM81 Condenser Handheld Microphone and Pop Filter, and with a sampling rate of 44.1 kHz. Each base recording was originally produced with relatively flat intonation. We then used PSOLA resynthesis in Praat (Moulines & Charpentier 1990; Boersma & Weenink 2020) to overlay new F0 contours on the base files. The F0 contours for resynthesis were based on the ToBI[1] training materials (Veilleux et al. 2006) which are in turn based on schema from Pierrehumbert (1980). The stimuli in each experiment were designed independently of one another, and thus made use of slightly different target values and landmarks for temporal alignment. The target values and corresponding F0 values for each speaker are shown in **Table 2**.

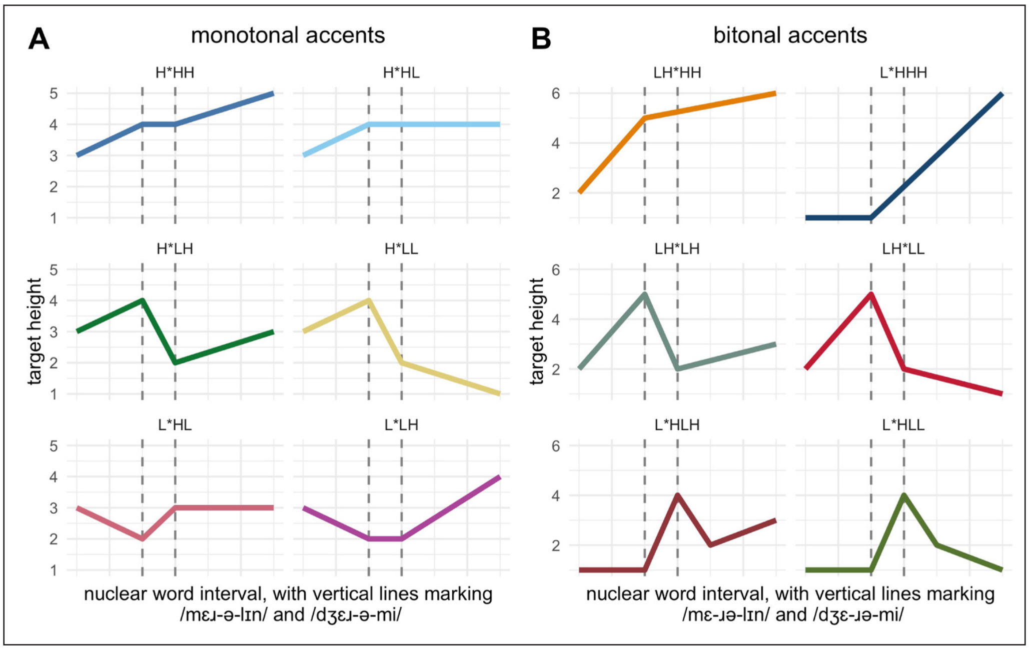


Figure 1: Schematic representation of the stimuli for the monotonal pitch accent experiment (Panel A), and the bitonal pitch accent experiment (Panel B). The y axis shows target values, which can be mapped to actual F0 values for each speaker by looking at their correspondence in Table 2. The x axis represents the interval of the nuclear word, the two dashed lines indicate temporal landmarks used for stimulus creation, which correspond to the locations in the segmental string indicated in the axis label by dashes (these differ slightly across experiments). The lines are spaced to reflect the average syllable duration across the model stimuli. See text for details.

[1] ToBI is an acronym for the Tones and Break Indices system, an annotation standard for intonation in the AM framework, and the tone features associated with each tune are referred to as its "ToBI label".

speaker measure		monotonal accents				bitonal accents			
		male		female		male		female	
		Hz	ERB	Hz	ERB	Hz	ERB	Hz	ERB
target	1	80	2.79	100	3.37	80	2.79	100	3.37
	2	105	3.51	160	4.93	105	3.51	160	4.93
	3	130	4.18	200	5.41	130	4.18	200	5.84
	4	225	6.36	300	7.79	225	6.36	300	7.79
	5	265	7.15	380	9.09	240	6.67	350	8.62
	6	—	—	—	—	260	7.15	380	9.09

Table 2: The values used for target heights in both experiments, numbered from 1–6 and shown in Hz and ERB. Note that the monotonal accent experiment used only five target heights.

The goal in creating the stimuli was to make a set of tunes that were perceptually different from one another and accurately reflected how tunes are schematized in Veilleux et al. (2006), while also being tightly controlled in terms of resynthesis. Three ToBI-trained experts (two of whom were the authors) judged the tunes to be distinct from one another within an experiment, and an appropriate realization of each of the model-defined tunes. All stimulus files can be found online on the open access repository.

To describe the model stimuli we will make use of two terms: *temporal landmarks* and *targets*. Temporal landmarks define when (in time) in the nuclear word F0 can change or reach a particular value. Interpolation between the landmarks is linear. Targets are F0 values, which we map to abstract numerical target height numbers to generalize across the two model speakers for the purpose of creating appropriately comparable F0 patterns that still fall within each speaker's F0 range. In other words, target heights are relative F0 within a speaker's F0 range. For example, for the male model speaker, target height 1 (the lowest) corresponds to 80 Hz. For the female speaker it corresponds to 100 Hz. This mapping was determined by the above-mentioned auditory assessment of the naturalness of the tunes by three ToBI-trained phoneticians (two of whom were the authors). **Table 2** shows the correspondence between abstract target values and actual F0 values for both model speakers and in both experiments. Temporal landmarks for the placement of targets were defined on the basis of segmental landmarks as assessed auditorily and by examination of the spectral transitions between segments. The chosen landmarks were slightly different between experiments, which we describe in turn.

In the monotonal accents experiment, the first temporal landmark corresponded to the point in time at the end of the third segment in the nuclear word, which was the onset of the second syllable. The second temporal landmark corresponded to the onset of the final syllable. Indicated with dashes for both nuclear words this is: /dʒɛɹ-ə-mi/ and /mɛɹ-ə-lɪn/. **Figure 1** provides a schematic representation of the tunes, using target heights on the y axis, and an abstract representation of time on the x axis. The dashed vertical lines mark the times corresponding to the landmarks, and they are spaced to reflect the average syllable duration across the model stimuli. Panel A can therefore be read as follows, for the tune H*HH (for example), at the onset of the nuclear word F0 is at target height 3 (male speaker: 130 Hz, female speaker: 200 Hz). At the time in each of the four base files corresponding to the first temporal landmark (first dashed vertical line) F0 has interpolated linearly up to target height 4 (male speaker: 225 Hz, female speaker: 300 Hz).

The principle for stimulus creation is exactly the same in the bitonal accent experiment, for which the stimuli are shown schematically in **Figure 1**, Panel B. However here, it was determined that the tunes in that experiment sounded more natural with a different organisation of temporal landmarks. The first landmark was moved slightly earlier in time, now corresponding to the boundary between the first and second syllable: /dʒɛ-ɹə-mi/ and /mɛ-ɹə-lɪn/. In order to produce the correct trajectory for two tunes, L*HLH and L*HLL, one additional F0 turning point was needed (see **Figure 1**, Panel B). The temporal landmark for this turning point was set to be one third through the duration of the final syllable. Note that the bitonal accents experiment additionally makes use of one additional target height, and the mapping to F0 values is slightly different, as shown in **Table 2**.[2]

The set of six tunes in each experiment were developed independently from one another. It can be noted that the implementation of the L* target in the monotonal versus bitonal experiment differs. In the former, L*LH and L*HL show F0 falling through the first (stressed) syllable, while f0 is level and lower in the bitonal accent experiment. The falling F0 in the monotonal L* was judged to sound most natural, given the tune as a whole and preceding pre-amble. It should thus be kept in mind that the shape of the L* target of the monotonal L* and the bitonal LH*, L*H pitch accents are not directly comparable across experiments.

[2] The F0 on the preceding preamble in the monotonal accents experiment was the same across the six tunes selected as stimuli, as each nuclear tune started from the same value. The preamble started at the value of target level 3 plus 20Hz, stayed at this value for the first third of the preamble, and then fell linearly to the value of target level 3 at the start of the nuclear word. In the bitonal accents experiment, it was deemed necessary to vary the preamble as a function of pitch accent, because the two pitch accents started at different target levels. For tunes containing L + H*, the starting point in the preamble was the same value as target level 3 plus 20Hz, with a linear fall directly to target level 2 at the start of the nuclear word. The starting point was the same for L* + H, but the fall was to target level 1 at the start of the nuclear word.

2.3 Participants and procedure

We recruited 32 participants for each experiment, all of whom were self-reported monolingual speakers of American English. No participant took part in both experiments. Participants were recruited from the undergraduate population of Northwestern University, and received course credit for their participation in the experiments. The studies were granted ethics approval by the university's institutional review board. In the monotonal accent experiment, participants had a mean age of 19.8 (range: 18–21 years) and a gender breakdown of 16 women, 15 men, and one non-binary participant. In the bitonal accent experiment, participants had a mean age of 19.6 (range: 18–22 years) and a gender breakdown of 15 women, 16 men, and one non-binary participant. Participants' regional background ranged across the United states; including the west coast, south and east coast. Many of the participants were from the midwest: 15/32 in Experiment 1 and 11/32 in Experiment 2. We are not in a position to assess regional, dialectal, or sociolectal variation in this study, though this would certainly be a valuable extension, which we discuss in Section 4.2.

Participants provided informed consent to participate in the study and received course credit for their participation. Participants completed the study remotely, using their own computer for the display of the experiment and audio recording. Each experiment took approximately 20–25 minutes to complete. The software used to present stimuli and record audio from the participants was a custom-built web-based application.

There were a total of 120 trials which crossed the six tunes in each experiment with the ten target sentences and the order of model speakers (male speaker first versus female speaker first) and model sentences ("He answered Jeremy" first versus "Her name is Marilyn" first). To limit the number of trials we did not fully cross the model speakers and model sentences with tune and metrical condition (which would have produced 240 trials). Instead, we paired each sentence order and each speaker order an equal number of times with each metrical condition (though not with each unique target sentence), also ensuring that each model speaker order and each sentence order appeared an equal number of times during the experiment. This results in the total breakdown of trials as: 6 tunes × 10 target sentences × 2 speaker/sentence levels.

2.4 Measurement and analyses

In this study we focus on only the nuclear tune, and we do not analyze the material that preceded the nuclear tune. We measured F0 using the STRAIGHT algorithm in Voice-Sauce (Kawahara et al. 2005; Shue et al. 2009), which we set to take measurements at every 10 ms. We then used the Montreal Forced Aligner (McAuliffe et al. 2017) to force-align text grids which segmented sound files by word and by phone. Forced-alignment was manually checked by trained auditors to ensure the start and end of the nuclear word was accurately aligned. Within the nuclear word, phone boundaries corresponding to syllable boundaries were also checked, and corrected to be anchored to acoustic landmarks when needed. During the auditing process, files that contained unusable audio quality, hesitations and speech errors were excluded (approximately 2.7% in the monotonal accent experiment, 2.6% in the bitonal accent experiment).

Before analyzing the data, we eliminated unreliable F0 measurements from our analysis. This is an issue, particularly in our case, because F0 measurement can be disrupted phrase- and utterance-finally due to non-modal phonation or low amplitude voicing. Because we are analyzing F0 at the end of an utterance, we wanted to be sure our measures were reliable. We did this using a semi-automated method that involved two steps. First, using the algorithm described in Steffman & Cole (2022), we identified F0 trajectories that contained a sample-to-sample change which exceeded previously described thresholds of F0 changes produced by speakers, using the thresholds documented in Sundberg (1973). This method effectively identifies sudden jumps in F0 measurement which are likely to be inaccurate measurements. These flagged files were then inspected by a trained auditor who either confirmed or disconfirmed an F0 measurement error (see Steffman & Cole 2022 for details and examples). In total, this method led to the exclusion of approximately 11% of the remaining files in both experiments.[3] In total, we analyzed 3,340 productions from the monotonal accent experiment and 3,330 productions from the bitonal accent experiment.

We extracted various parameters from the measured F0 over the nuclear word, which are described below. All of the analyses were carried out using R (version 4.1.2) in the RStudio environment (R Core Team 2021; Posit team 2023).

2.4.1 The logic of the analyses

We pursued several analyses which were intended to complement eachother and address the question of enhancement. In this vein, it is important for us to operationalize enhancement, and define what would constitute evidence for it. In this section we thus motivate the analyses and explain how they test for enhancement.

As one of our principle goals was to determine the distinctions among tunes that are most readily produced by speakers, we carried out a clustering analysis on data that was not labeled by tune. This is a “bottom up” analysis in the sense that the clustering algorithm does not have access to the pre-conceived tune category labels. Our main interest in the cluster analysis was the *number of clusters* that best partitioned the data, and the mapping between model tunes

[3] This method of F0 error tracking did not impact each tune equally. As might be expected, lower and falling boundary tones were excluded the most in both experiments LL-ending tunes were impacted the most (80% of files retained as compared to 94% for the most retained files in LH*HH in Experiment 2 (cf. 90% in H*HH in Experiment 1). We note that importantly, data retention does not correlate with clustering separability, as will be seen in the results. Though manual correction of F0 would be possible in principle, in almost all cases here the signal was distorted to the extent that it was not clear was actual F0 values were, making manual correction tenuous.

used as stimuli (six in each experiment) to clusters. This effectively allows us to examine if distinctions among all six model tunes are robust (i.e. six clusters corresponding to the six tunes), and if distinctions are enhanced or lost across metrical conditions. This is the first way that we consider enhancement: the number of clusters in the optimal clustering solution increases correspondingly with increases in metrical material (from metrical condition 1 to 4b). If observed, this would constitute evidence that enhancement is present in the emergent distinctions speakers produce. In other words, speakers make more and/or better-separated distinctions in F0 when the phonological material carrying the tune is longer. Relating these patterns to the tune labels that correspond to the clusters will further allow us to see if, for example, two tunes that cluster together (suggesting a lack of distinctiveness), fall into separate clusters when there is more metrical material.

The second way we consider enhancement is in acoustic (F0) space. In other words, do tunes become measurably more different from one another in some parameter with increasing metrical material? We make use of three different F0 measures, which are described in Sections 2.4.3 and 2.4.4. If a particular difference between tunes becomes larger with increasing metrical material, this constitutes evidence for enhancement. Here too we also make reference to the clustering analysis, in considering if enhancement is observed between MAE-labeled tunes that are separated in the clustering solution vs. those that are merged. The purpose of this evaluation is to ask if and how clustering predicts enhancement: if distinctions between tunes that cluster-together are enhanced, this suggests the clustering analysis is missing subtler enhancement effects. If distinctions between tunes that cluster separately are enhanced, this suggests that the emergent cluster categories are targets for enhancement, i.e., that being partitioned into separate clusters predicts that two tunes will become more separated with increasing metrical material.

Given the multi-step analysis plan and rather large set of comparisons and results, we conclude each sub-section in Section 3 with a brief synopsis that describes how that section fits into the larger question of enhancement and the analysis as a whole.

2.4.2 Time series clustering analysis

We used the *kml* package (Genolini & Falissard 2011) to carry out k-means clustering analyses on F0 time series data, separately for the monotonal and bitonal datasets. We take two approaches to clustering the data in each experiment. In one, we compute speaker means by tune while averaging across metrical conditions (6 total trajectories per speaker, one for each tune). This clustering solution effectively captures the best overall partition of the data independent of metrically-induced variation. This approach further ensures that each speaker contributes equally to the data and also allows us to average over some trial-to-trial variation in the speakers' productions. Secondly, we compute speaker means by tune within each metrical condition, comparing clustering solutions across metrical conditions (30 total trajectories per speaker: 6

tunes × 5 metrical conditions). This approach allows us to examine how emergent distinctions in the clustering solution may change, and in particular, if an expanded metrical structure leads to more, or better separated, distinctions. Both of these approaches are presented in the results.

We computed F0 using ERB, and further scaled the measures within each speaker (z-scored, with reference to each speaker's mean F0 value). This scaling effectively normalizes for between-speaker differences in overall F0 height and F0 range. F0 was measured at 30 time-normalized samples over the interval of each nuclear word. The optimal partition of the data into clusters was determined using the Caliński Harabasz criterion (Caliński & Harabasz 1974), which tests various numbers of cluster as potential solutions and identifies as optimal the number of clusters that minimizes within-cluster variance while maximizing between-cluster variance (computed using Euclidean distances between trajectories). We consider solutions with as few as two and as many as ten clusters.

2.4.3 Analysis of Root Mean Squared Distance (RMSD)

In order to examine enhancement in F0 space we computed Root Mean Squared Distance (RMSD). This metric, described below, is a holistic measure of F0 space/difference between two trajectories, which is global in the sense that it is not localized in time. For this reason, it is well suited to comparing tunes which are very different from one another (which will have larger RMSD), as well as those that are more similar (which will have small RMSD).

RMSD was computed over the time-normalized F0 trajectories comparing by-speaker and by-metrical structure means for each tune (in ERB). In other words, each speaker's mean F0 trajectory for a production of a given tune, in a given metrical condition was computed. This totalled 30 trajectories per speaker: six for each tune (separately for the monotonal and bitonal datasets), within each of the five metrical conditions. Then, within a given metrical condition (and separately for each of the two datasets), all tunes were paired with one another in all possible combinations, yielding 15 tune pairs (all possible order-insensitive pairings of the six tunes). The RMSD formula in (1) was then applied to each pair of tunes, where x is one tune, y is the other, and N = 30, for each of the 30 time points in the time-normalized trajectories.

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

This metric effectively captures the overall distance or separation between a particular pair of tunes in F0 space for a given speaker, and because it was computed for each metrical condition, allows us to see if the separation between tunes, as measured with RMSD, increases with increases in the metrical material to carry the tune. This measure will also be considered with respect to variables derived from the clustering analysis, described in Section 3.2.

To model these influences on RMSD we use Bayesian linear mixed effects models, implemented with the *brms* package in R Bürkner (2017).[4] The dependent variable was RMSD in ERB. When including the metrical structure variable in these and subsequent models, we treated it as a monotonic effect (Bürkner & Charpentier 2020): an ordinal predictor which assumes changes across metrical conditions will be monotonic, while also allowing for (potentially) different effect sizes across levels of the metrical condition.

In these models, and others below, we report effects in several ways. We first give the median posterior estimate ($\hat{\beta}$) and the 95% credible intervals (CrI) for that estimate. These intervals indicate the range in which 95% of the estimates for an effect fall. When the interval excludes the value of zero, this is taken as credible evidence that the effect is non-zero, with a particular directionality (conversely, intervals including zero indicate substantial variation in the estimated directionality of an effect). We also report the "probability of direction" metric, which is computed from an effect's distribution and gives the percentage of that distribution which shows a particular directionality. A pd of 97.5% corresponds to 95% CrI excluding zero, and pd can be interpreted more intuitively as strength of evidence for an effect, i.e. if pd = 95 we can be 95% sure that an effect exists with a particular directionality. In some cases we also report estimates and pd for marginal effects, which we compute using the package *emmeans* (Lenth 2021).

2.4.4 Other analyses: tune-based comparisons across metrical structures

We carried out several additional analyses to compare F0 trajectories between pairs of tunes that are similar to each other and which tend to cluster together in the cluster analyses. The motivation for these particular comparisons will be described in more detail below.

To assess the presence and enhancement of distinctions corresponding to boundary tones in the tune labels we took F0 at the end of the nuclear word, and tested, for a given pair of tunes, if there was a detectable difference and if this changed as a function of metrical structure. These models were each carried out with a pair of tunes (described in Section 3.3). Given that the variable had only two levels, we contrast-coded it (e.g., tune 1 mapped to –0.5, and tune 2 mapped to 0.5). Each model predicted ending F0 as a function of tune pair, metrical structure, and their interaction. Random effects included by-speaker intercepts with random slopes for both fixed effects and their interaction. More specific predictions will be discussed in Section 3.3.

We also examined tonal timing to provide a window into another potentially important cue to distinctions in the nuclear tunes. In lieu of using more traditional measures of timing such as peak or valley alignment, we computed the temporal Tonal Center of Gravity (TCoG) over

[4] For all mixed-effects models reported here, the model drew 4,000 samples in each of four Markov chains, with a burn-in period of 1,000 iterations (75% of samples retained for inference). The *adapt delta* parameter was set to be 0.99 in all models. $\hat{R}$, Bulk ESS, and Tail ESS values were examined to confirm convergence and adequate sampling.

the nuclear tune. TCoG has been proposed as a holistic measure to model tonal timing for pitch accents (Barnes et al. 2012; 2021), and captures when in time the bulk or mass of a higher-F0 region occurs. This is computed using the equation shown in (2), where $F0_i$ is F0 at time i, t_i is that time in milliseconds, and N is the total number of samples in the window, which we measured in 10 ms intervals.

(2)
$$temporal\ TCoG = \frac{\Sigma_{i=1}^{N} F0_i \times t_i}{\Sigma_{i=1}^{N} F0_i}$$

The intuition behind the above formula is that regions with higher F0 effectively pull TCoG towards them, and the formula integrates over a window such that rise and fall shape for a high F0 region contribute to the computation. The reader is referred to Barnes et al. (2012; 2021), for further discussion of TCoG. In a departure from previous work using TCoG, we apply the measure to the nuclear region as a whole for all metrical conditions, thus computing TCoG over up to four syllables. Whereas the measure has previously been applied to rising-falling F0 movements that constitute a "hump" in the F0 contour, we also compute TCoG for monotonically rising F0 movements. Just as with a rising-falling movement, rising F0 can be described in terms of TCoG. This approach is taken to model two pairs of tunes whose F0 trajectories have an overall rising shape: {H*HH, H*HL} from the monotonal experiment, and {L*HHH, LH*HH} from the bitonal accent experiment. We also use TCoG to describe the placement of the higher F0 region in trajectories of the tunes {L*HL, L*LH} from the monotonal accent experiment, which have an overall dipping and falling-rising shape. Finally, in a more traditional application of the measure, we compare the TCoG for two rising-falling tunes in the bitonal accent experiment: {L*HLL, LH*LL}. The choice of each of these tune pairs for comparison will be motivated in the results section. Note too that we do not apply this analysis to H*LH (from the monotonal accent experiment), L*HLH or LH*LH (from the bitonal accent experiment). Each of these tunes is predicted to have multiple F0 "humps", which would make TCoG measures for them difficult to interpret.

To illustrate the general application of TCoG for this purpose, we plot each of the above-referenced tune pairs in **Figure 2**, showing the grand mean F0 (scaled) for each tune averaged across metrical conditions and speakers' productions in the experiments, and with a time-normalized representation of mean TCoG for each tune (also averaged across metrical conditions). Note that, as shown in Panel A of **Figure 2**, the two rising tunes {H*HH, H*HL} differ in their mean TCoG with H*HH showing a more scooped rise, and higher-ending F0 that effectively pulls TCoG later. The same can be said for the two rising pairs in the bitonal accent experiment, with L*HHH showing a more scooped rise, and later TCoG. {L*HL, L*LH} show a similar TCoG difference, wherein higher F0 at the end of the tune in L*LH pulls TCoG later. In the rising-falling pitch accent pair {L*HLL, LH*LL} in Panel B, the later rise of L*HLL pulls TCoG later.

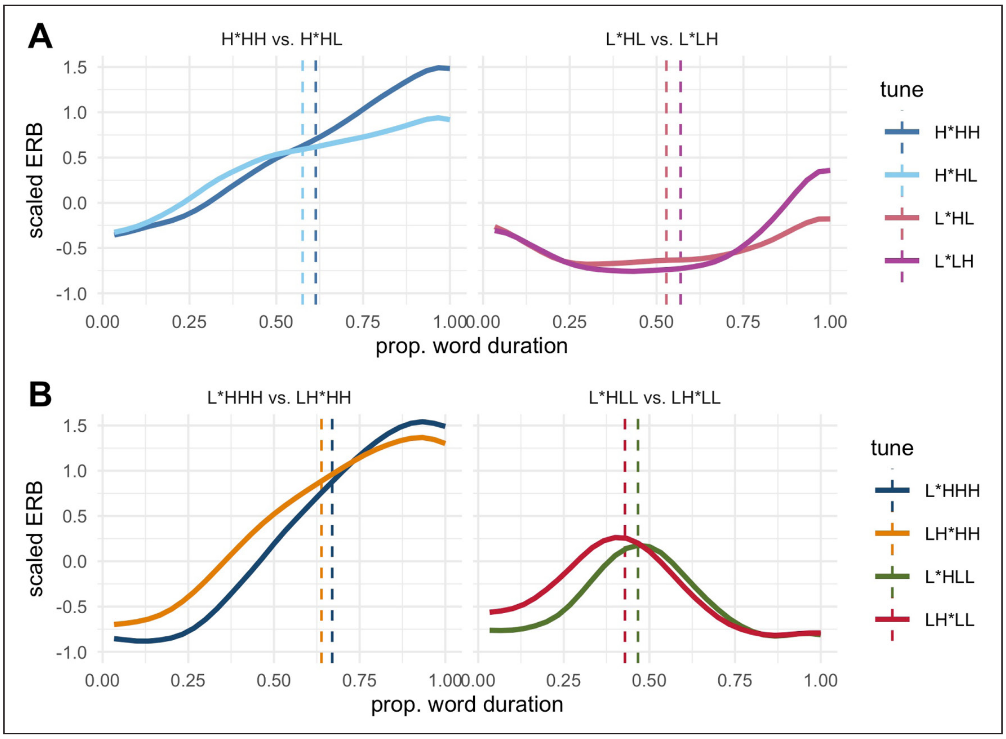


Figure 2: Mean trajectories and temporal TCoG (dashed vertical line), represented in normalized time as a proportion of nuclear word duration, for two pairs of tunes in the monotonal accent experiment (Panel A), and the bitonal accent experiment (Panel B). See text for details.

In the actual TCoG measurements we report in the results, we computed TCoG in raw time, with time starting at zero at word onset. This TCoG measure was anchored with respect to the end of the first syllable in the nuclear word (meaning that we did not carryout this analysis for one syllable words). The anchoring procedure was as follows: for a given production, the time (in ms) that the second syllable in the nuclear word starts (a boundary that was manually checked during text grid auditing) was subtracted from the TCoG measure. This measure is then effectively one of TCoG alignment. The anchoring point is also phonologically meaningful in that it is located at the boundary of the syllable that is pitch-accented and the following syllable. More practically, this TCoG alignment measure will indicate to us whether the TCoG is within the first syllable (negative values) or after the first syllable (positive values). As noted, given this method of computing TCoG, we do not analyze TCoG for tunes in the 1 syllable metrical condition, for which the distance from the end of the first and only syllable will always be negative. We also do not analyze TCoG in the 4b metrical condition, since we expected that secondary stress could potentially result in multiple F0 masses over the tune, which would make TCoG difficult to interpret.

The OSF repository for the paper contains an additional analysis using Generalized Additive Mixed Modeling (GAMMs) to model differences between certain tune pairs over time. We opted not to include this analysis here since it largely confirms our conclusions from the other analyses. The interested reader is referred to the document `GAMM_supplement.pdf` on the online repository which provides the methods and summarizes the results from the GAMM analysis (most easily read after reading this paper).

3 Results

In this first section of results we simply take a birds-eye view of the trajectories across tunes and metrical conditions, as well as considering how the duration of the nuclear interval varies across both of these factors. **Figures 3** and **4** show all of the trajectories analyzed in each experiment. In Panel A in each figure, tunes are shown in columns and metrical conditions in rows. Within each panel, all of the trajectories are shown over time (in milliseconds) and represented as scaled ERB, normalizing for differences in speaker F0 levels and F0 range. The dashed vertical line represents the mean duration of trajectories in that cell. As can be seen looking across the rows, there is an overall increase in duration as the number of syllables increases across metrical conditions. Panel B in each Figure shows the duration of each tune in each metrical condition as a violin plot. For both the monotonal and bitonal pitch accent experiments we note the following: The general shape of trajectories for each tune does not change dramatically in longer words, though some smaller modulations occur; H*HH for example seems to have a steeper (more-scooped) initial rise overall in 1,2, and 3 syllable metrical conditions. Importantly, the clear (and expected) increase in duration suggests that there is at least a possibility for the enhancement of tunes across metrical structures of increasing length and complexity.

In a model of word duration predicting the (logged) duration of the nuclear word as a function of tune, metrical structure and their interaction (with by-speaker random intercepts and fully specified random slopes), we extracted marginal effect estimates using *emmeans* to compare across levels of the metrical structure variable. We found that there were indeed credible differences across all levels such that 1 < 2 < 3 < 4a < 4b (all pds = 100 for both the monotonal and bitonal experiments. In the monotonal experiments the following marginal means (back transformed from logged values) were 387, 468, 542, 576, and 663 ms across metrical conditions (1-4b). The bitonal accent experiment, with data shown in **Figure 4**, showed the following estimated marginal means: 424, 503, 572, 615, and 692 ms across metrical conditions (1-4b).[5] Overall, our first view of the data suggests that, unsurprisingly, words with more syllables are longer, with additional length in the four-syllables words with secondary stress, and also, qualitatively speaking, the overall shape of trajectories for a given tune does not change dramatically across metrical structures.

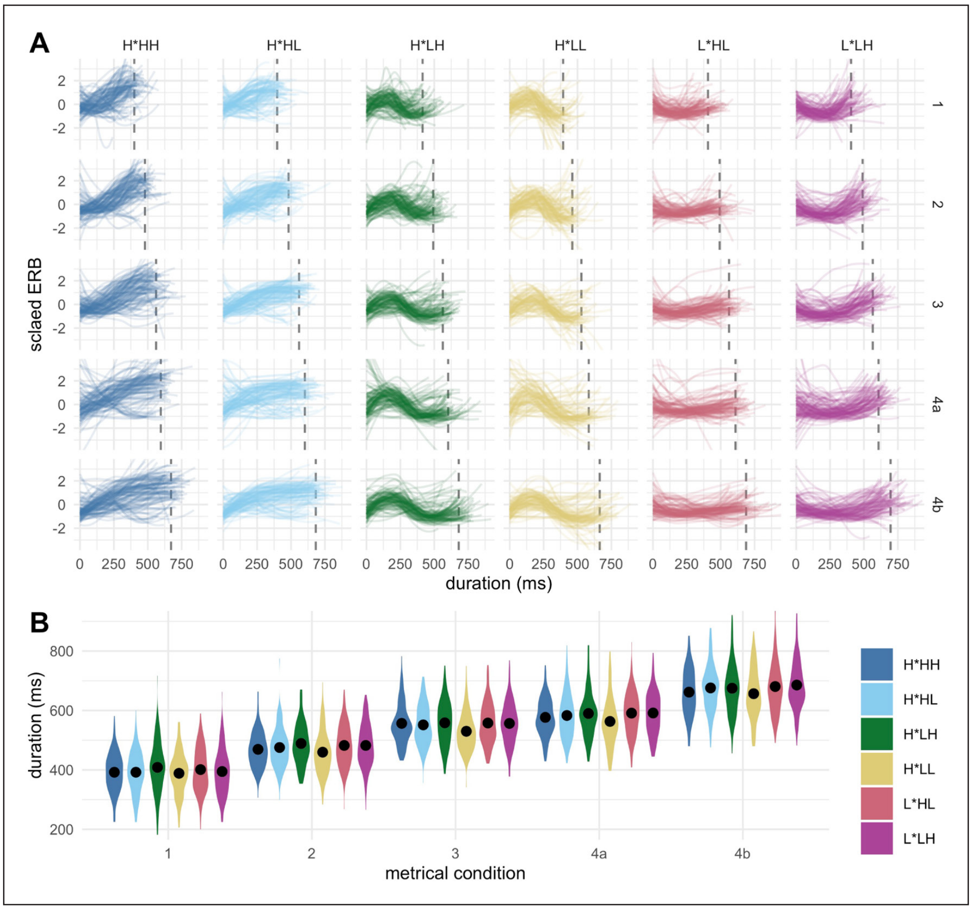


Figure 3: Panel A: Trajectories across tunes (columns) and metrical conditions (rows), in the monotonal pitch accent experiment. The duration of the nuclear word is plotted on the x axis, and the dashed vertical line shows the mean nuclear word duration for trajectories within each sub-panel. Panel B: violin plots showing nuclear word duration across metrical conditions (x axis) and tune (coloration), points show the mean.

[5] There were also some credible differences between tunes, though these were much smaller than the effects of metrical structure, the estimated marginal means for tunes (back-transformed to ms from logged values) were the following. In the monotonal experiment: H*HH = 515 ms, H*HL = 519, H*LH = 525, H*LL = 504, L*HL = 524, L*LH = 525. In the bitonal accent experiment: LH*HH = 552, LH*LH = 549, LH*LL = 545, L*HHH = 555, L*HLH = 560, L*HLL = 556.

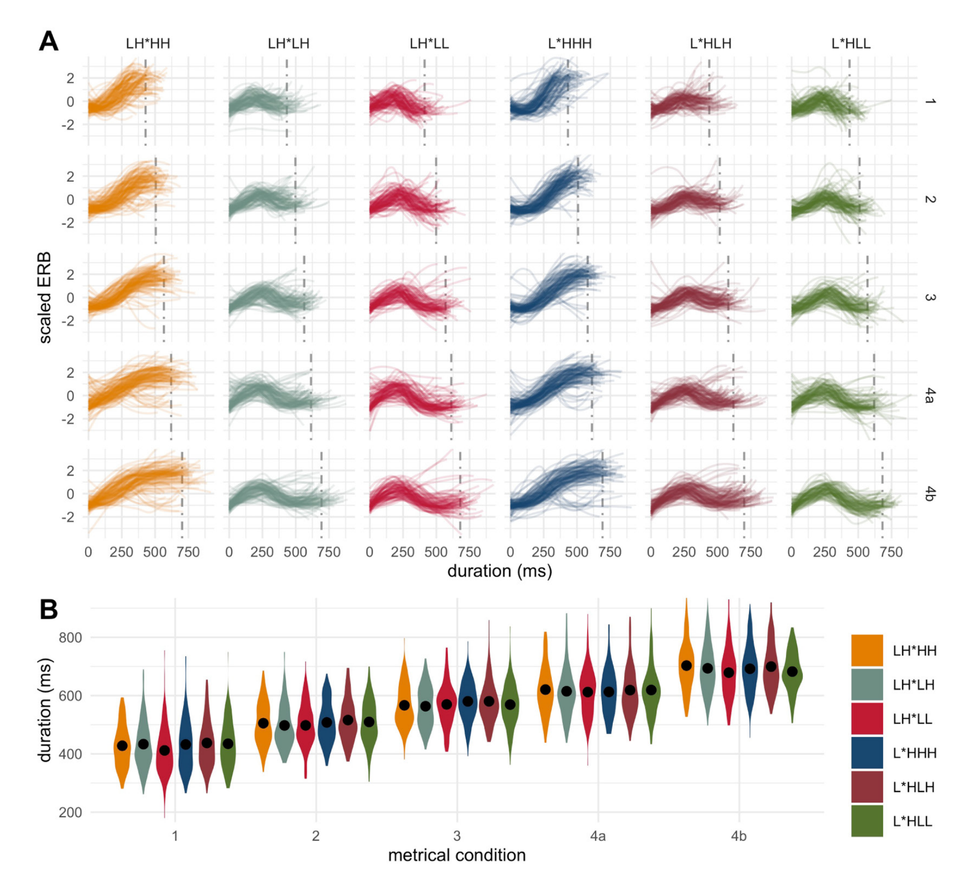


Figure 4: Panel A: Trajectories across tunes (columns) and metrical conditions (rows), in the bitonal pitch accent experiment. The duration of the nuclear word is plotted on the x axis, and the dashed vertical line shows the mean nuclear word duration for trajectories within each sub-panel. Panel B: violin plots showing nuclear word duration across metrical conditions (x axis) and tune (coloration), points show the mean.

3.1 Clustering results

Given the possibility for metrical enhancement suggested by the duration results above, this section examines the emergent distinctions in the data using clustering analysis. **Figure 5** shows the optimal clustering partition of the data in the monotonal accent experiment. This is shown in two ways: clustering over speaker-mean trajectories aggregated across metrical structures (Panel A), and clustering over speaker means within metrical structure (Panel B). Our key interest in analyzing the clustering results was to determine the optimal number of clusters returned by the algorithm and the mapping of tunes to clusters to assess which tunes reliably cluster together (or apart). As shown in **Figure 5** Panel A, the aggregated means are optimally partitioned into three clusters, each of which is primarily composed of two tunes: L*HL and L*LH in cluster A, which shows a low-rising shape; H*LL and H*LH in cluster B, which shows a rising-falling shape; and

H*HL and H*HH in cluster C, which show high-rising shape. The fact that only three clusters are emergent with six tunes in the input suggests that distinctions between tunes grouped in the same cluster are small and inconsistent and support only a non-optimal partition of the data, i.e. poorly-separated and diffuse clusters. As shown in Panel 5B, cluster solutions for the 3, 4a and 4b metrical conditions show a highly comparable partition of the data: three clusters with the same merged tune pairs. Taking the 3-syllable metrical condition as a baseline (matching the metrical pattern of the stimulus), these results indicate an absence of enhancement in that the longer (4-syllable) metrical conditions result in no additional clusters beyond the three that emerge from the 3-syllable condition. Conversely, the 1 and 2 syllable metrical conditions suggest a loss of distinctions, showing only two clusters as optimal, with L*HL, L*LH, H*LH, and H*LL clustering together into a single "non-high-rising" cluster. Distinctions between any pair of tunes in this group have been reduced such that they are not large or systematic enough to lead to distinct cluster partitions.

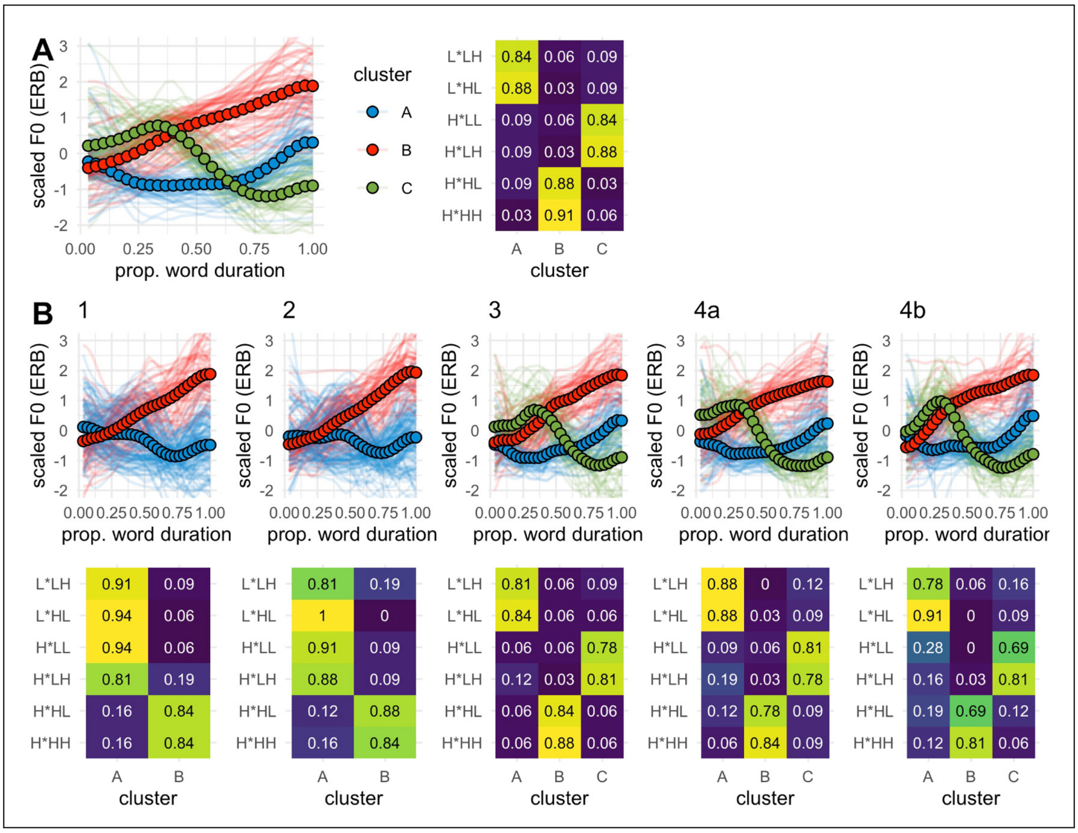
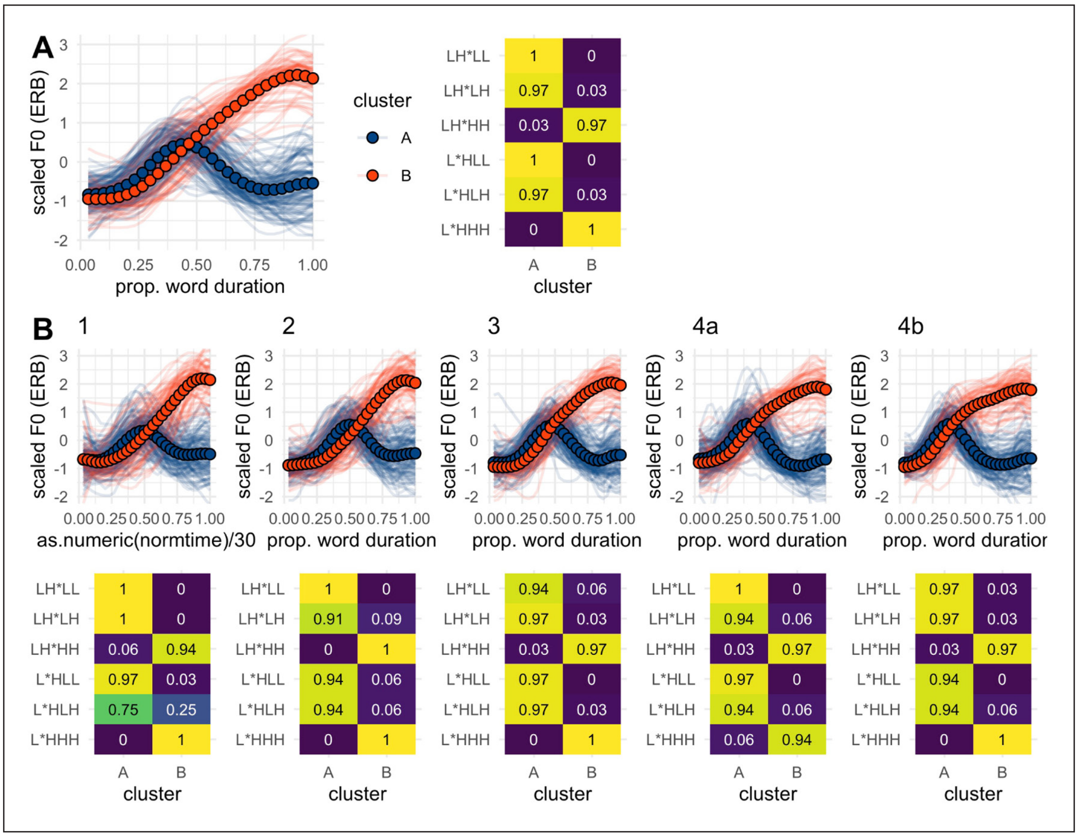


Figure 5: Panel A: The optimal clustering solution for the monotonal accent experiment. The trajectory plot shows cluster means as points, and all contributing trajectories as lighter lines. The heat maps indicates the proportion of each tune (rows) that went into each cluster (columns). Panel B shows the same information, with clustering carried out within each metrical condition.

Figure 6 shows the results of cluster analysis for data from the bitonal experiment in the same way. Here, unlike with the monotonal pitch accent experiment, the optimal clustering partition for the aggregated trajectories, and also for each metrical condition, is two clusters. These two clusters group together the two tunes with HH edge tones (LH*HH and L*HHH), into a single rising cluster, suggesting the two pitch accents are not well distinguished. The remaining four tunes (L*H and LH* concatenated with both the LL and LH edge tones) cluster together in a single rising-falling cluster. The similarity of the aggregated cluster analysis with each of the metrical conditions, and the lack of change across metrical conditions shows that this rising versus rising-falling distinction is quite robust. Conversely, differences among the four tunes grouped together in the same cluster are again insufficient to support additional clusters. With the 3-syllable metrical condition as a baseline, we again conclude that these small differences are not enhanced with an expanded metrical structure.



A:

	A	B
LH*LL	1	0
LH*LH	0.97	0.03
LH*HH	0.03	0.97
L*HLL	1	0
L*HLH	0.97	0.03
L*HHH	0	1

B, 1:

	A	B
LH*LL	1	0
LH*LH	1	0
LH*HH	0.06	0.94
L*HLL	0.97	0.03
L*HLH	0.75	0.25
L*HHH	0	1

2:

	A	B
LH*LL	1	0
LH*LH	0.91	0.09
LH*HH	0	1
L*HLL	0.94	0.06
L*HLH	0.94	0.06
L*HHH	0	1

3:

	A	B
LH*LL	0.94	0.06
LH*LH	0.97	0.03
LH*HH	0.03	0.97
L*HLL	0.97	0
L*HLH	0.97	0.03
L*HHH	0	1

4a:

	A	B
LH*LL	1	0
LH*LH	0.94	0.06
LH*HH	0.03	0.97
L*HLL	0.97	0
L*HLH	0.94	0.06
L*HHH	0.06	0.94

4b:

	A	B
LH*LL	0.97	0.03
LH*LH	0.97	0.03
LH*HH	0.03	0.97
L*HLL	0.94	0
L*HLH	0.94	0.06
L*HHH	0	1

cluster

Figure 6: Panel A: The optimal clustering solution for the bitonal accent experiment. The trajectory plot shows cluster means as points, and all contributing trajectories as lighter lines. The heat maps indicates the proportion of each tune (rows) that went into each cluster (columns). Panel B shows the same information, with clustering carried out within each metrical condition.

Synthesising the two cluster analyses, we make the following observations. First, emergent clusters (in both the aggregate and metrical-condition-specific analyses) collapse some tunes:

many of the predicted tune distinctions are not emergent in a bottom-up partition of the data. Note that this does not necessarily mean that there are *no* differences between two tunes that cluster together, but rather that distinctions, if present, are small and noisy to the extent that they do not justify an additional partition of the data into distinct clusters in the optimization component of the clustering analysis. This is explored in subsequent analyses below. Evidence for enhancement across the metrical conditions is also limited. In the monotonal accents experiment, some evidence for a loss of distinctions from the baseline 3 syllable condition to the 1 and 2 syllable conditions is present. However, the lack addition of clusters in the 4a and 4b conditions speaks against enhancement from the baseline. The highly-similar partitions of the data across metrical conditions in the bitonal accents experiment further comports with this.

3.2 RMSD analysis

The clustering analysis defines groups of tunes that are well distinguished from one another, and groups of tunes that are not. In this section we examine how this emergent, data-driven, partition may be related to enhancement, as outlined in Section 2.4.1. We use the term "between-cluster" for a pair of tunes whose imitations mostly fall into different clusters and "within-cluster" for a pair of tunes whose imitations mostly belong to the same cluster. Our interest in the RMSD analysis was to consider how the distance in F0 space between a pair of tunes varies as a function of metrical structure, and whether this relationship may be mediated by their status as between- or within-cluster. In other words, this analysis asks if the RMSD for all tune pairs is enhanced across metrical conditions, or alternatively, if distinctions between only certain tune pairs are enhanced. The fact that additional clusters did not emerge in the 4a and 4b metrical conditions, relative to the 3-syllable condition, already suggests that indeed, enhancement for within-cluster tunes is limited.

If enhancement effects are observed only for between-cluster tune pairs, this would constitute a scenario in which "the rich get richer", that is, distinctions that are already robust (based on clustering) become larger across metrical conditions. To examine this, we structured our model to predict (scaled) RMSD as a function of tune pairs (recall that RMSD is computed on a by-tune-pair basis, with each speaker contributing 15 values per metrical condition, for 15 pairs). Two additional predictors in the model are metrical structure and a factor that we call "cluster class", which encodes whether a particular tune pair is between- or within-clusters. We also interacted these two fixed effects and structured the random effects in the model to include by-speaker random intercepts with both fixed effects and their interaction as random slopes. Given that there are only two levels, cluster class was contrast-coded with between-cluster mapped to –0.5, and within-cluster mapped to 0.5. The main point of interest will be in examining if there is an effect of metrical structure, i.e., enhancement of differences between tunes across metrical structures, and secondly, if this enhancement effect operates differently based on cluster class.

Cluster class was determined based on the aggregated clustering solution (with input trajectories averaged across metrical conditions; shown in **Figure 5** and **Figure 6** panel A). We took this clustering analysis to be most representative of the data as a whole. We additionally consider the overall propensity of a tune to go into a given cluster in making this distinction. The clustering partitions are clear-cut in this sense, especially in the aggregated solution: for each tune, over 80% of the imitations of the tune are put into a single cluster. For the monotonal accent experiment, we thus only have three within-cluster pairs: {L*LH, L*HL} in cluster A, {H*HH, H*HL} in cluster B and {H*LL, H*LH} in cluster C. The overall results for the monotonal pitch accent experiment are shown in **Figure 7**. Tune pairs are sorted from overall lowest to highest RMSD in Panel A, and it can be noted that the within-cluster pairs at left have the lowest RMSD values overall (this is expected given the way the clustering algorithm operates). Moreover, we can note qualitatively that many between-cluster tune pairs seem to evidence enhancement, that is, higher RMSD, across the ordered metrical structures.

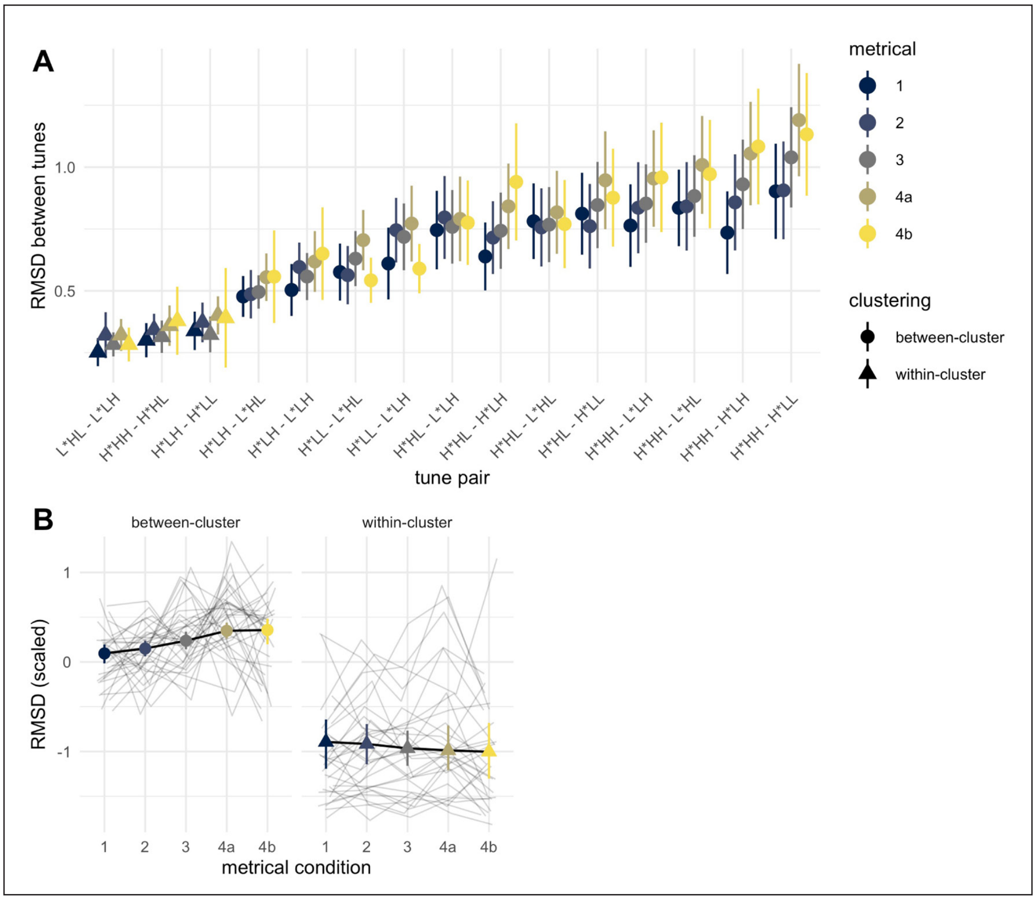


Figure 7: Panel A for each tune pair (x axis) across metrical conditions in the monotonal accent experiment. Point coloration indicates metrical condition, and the shape of the point indicates cluster class, based on the cluster partition in **Figure 5A**. The y axis is in un-transformed RMSD (ERB) values. Error bars show 95% CI computed from the raw data. Panel B shows model estimates in scaled ERB (points with 95%CrI from the model) based on cluster class (facets) and metrical condition (x axis). Lighter solid lines show individual participant data.

The modeling confirms this general observation, shown **Figure 7** Panel B.[6] Because the reference level in the model is by default the first level of the metrical condition, we computed marginal means to estimate the overall effect of cluster class, which was credible ($\hat{\beta}$ = 0.78, 95CrI = [0.61,0.95], pd = 100), showing that overall the between-cluster tune pairs have higher RMSD than within-cluster tune pairs. There was not a main effect of metrical structure ($\hat{\beta}$ = 0.01, 95CrI = [–0.03,0.05], pd = 65), however there was crucially an interaction between metrical condition and cluster class ($\hat{\beta}$ = –0.04, 95CrI = [–0.07,–0.01], pd = 99). The interaction was examined further by inspecting differences across metrical conditions within each cluster class. For the within-cluster tune pairs, there was no credible difference between any level of metrical condition, though there was weaker evidence of RMSD differences showing 2 > 3 ($\hat{\beta}$ = 0.03, 95CrI = [–0.00,0.07], pd = 96) and 2 > 4a ($\hat{\beta}$ = 0.05, 95CrI = [–0.00,0.10], pd = 97), though note that both of these weaker effects show the opposite directionality of enhancement (larger RMSD values in the 2-syllable metrical condition). In comparison, the between-cluster pairs showed a pattern that was consistent with enhancement, and several credible differences between adjacent pairs. The 1-syllable metrical condition was not credibly different than the 2-syllable condition, however the 2-syllable condition showed credibly smaller RMSD than the 3-syllable condition ($\hat{\beta}$ = –0.03, 95CrI = [–0.07, –0.00], pd = 99), and 3-syllable showed credibly smaller RMSD than the 4-syllable condition 4a ($\hat{\beta}$ = –0.02, 95CrI = [–0.05, –0.00], pd = 98). Metrical conditions 4a and 4b were not credibly different ($\hat{\beta}$ = –0.03, 95CrI = [–0.13, 0.08], pd = 72), though trended in the direction of enhancement. The interaction between cluster class and metrical structure thus reveals that, for between-cluster tune pairs only, RMSD increases in a way that is consistent with metrically conditioned enhancement. No such pattern is apparent for within-cluster pairs.

Figure 8 shows the data for the bitonal accents experiment. One important difference between the two experiments is that the number of within-cluster pairs in this experiment is greater (seven out of fifteen), though it can be noted that, as expected, all within-cluster pairs have smaller differences in RMSD than between-cluster pairs. The same qualitative pattern is apparent as well, overall larger differences are consistently observed for between-cluster pairs as a function of metrical structure, while this pattern is not readily apparent for within-cluster pairs.

The modeling results align with these observations and are similar to what was seen in the monotonal accents experiment. The marginal effect for cluster class showed a credible difference ($\hat{\beta}$ = 0.98, 95CrI = [0.81,1.14], pd = 100), whereby between-cluster tunes have higher RMSD, as expected. Unlike in the monotonal accent model there was an overall main effect of metrical structure ($\hat{\beta}$ = 0.05, 95CrI = [0.03,0.07], pd = 100), consistent with enhancement overall, though there was also critically an interaction with cluster class

[6] Note that, while **Figure 7** Panel B presents scaled estimates from a model run on scaled data for ease of visual inspection (normalizing over differences in speaker F0 range which lead to differences in RMSD), the model reported in the text was run on raw RMSD values, using a log-normal family to account for the all-positive and right tailed distribution of the RMSD data. Models run with non-logged values lead to very similar results. Both scaled and non-scaled models are also included on the OSF repository.

($\hat{\beta}$ = –0.05, 95CrI = [–0.09,–0.01], pd = 99), which indicates less of an enhancement effect for the within-cluster class, given its negative sign and the way that the variables were coded. Inspecting pairwise effects of metrical structure within cluster class confirms this. Within-cluster pairs show only one credible difference in adjacent metrical conditions, whereby $1 < 2$ ($\hat{\beta}$ = –0.06, 95CrI = [–0.12,–0.00], pd =99), with each other pairwise difference showing a non-credible effect (all pds < 87). In comparison, there is clear evidence for an enhancement effect in the between-cluster condition, whereby 1 < 2 < 3 < 4a < 4b (all pds = 100), as can be seen in Panel B of **Figure 8**. This inspection of the by-cluster-class effects of metrical structure thus suggests that the main effect of metrical structure is driven by the between-cluster pairs.

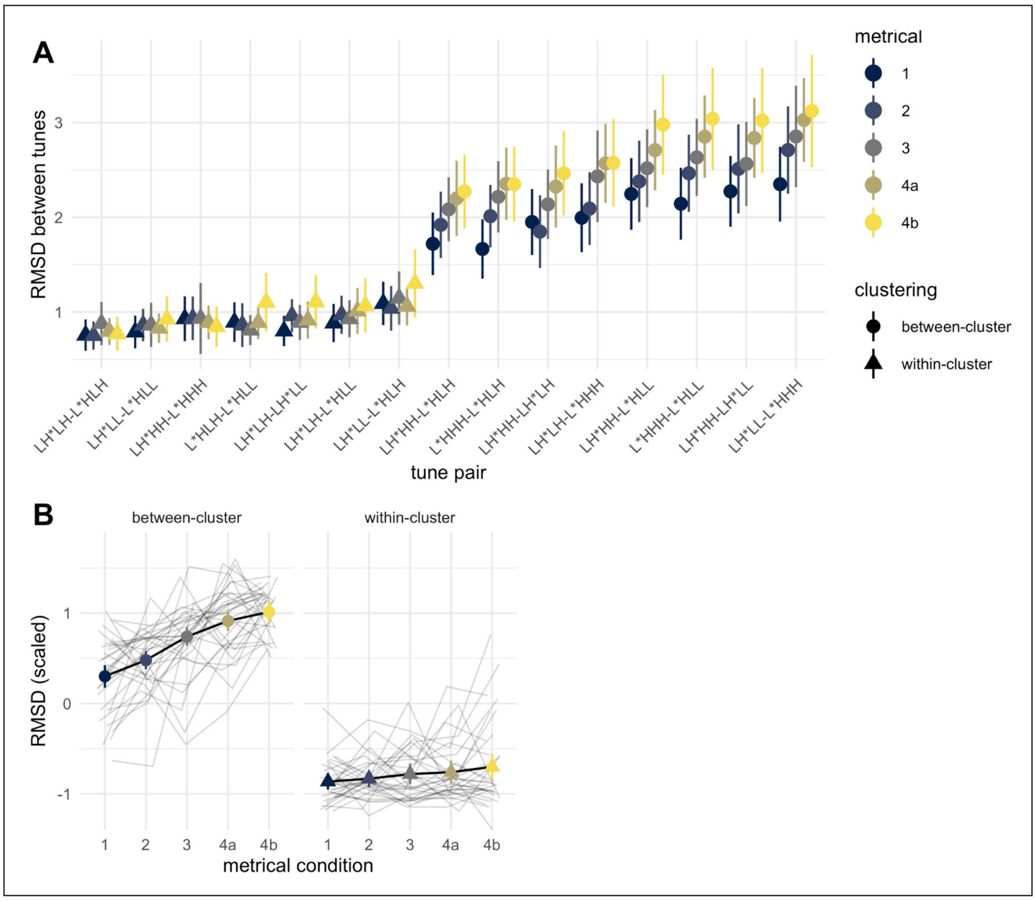


Figure 8: Panel A for each tune pair (x axis) across metrical conditions in the bitonal accent experiment. Point coloration indicates metrical condition, and the shape of the point indicates cluster class, based on the cluster partition in **Figure 6A**. The y axis is in un-transformed RMSD (ERB) values. Error bars show 95% CI computed from the raw data. Panel B shows model estimates in scaled ERB (facets) and metrical condition (x axis). Lighter solid lines show individual participant data.

Synthesising the two RMSD analyses, we find converging evidence that metrical enhancement targets tunes that are already distinct from one another (as operationalized by cluster class here). In other words, the largest and most salient distinctions in the data are enhanced as as a function of metrical structure, while smaller within-cluster distinctions are not. This result, perhaps surprisingly, suggests that, at least in terms of RMSD, there is not enhancement between same-clustering tunes. In the following section, we examine additional measures of F0 differences between tune pairs that clustered together, in order to confirm this conclusion and test for enhancement in a more local and targeted manner.

3.3 Ending F0 and TCoG

The preceding analyses have led to two somewhat striking results. First, not all tunes in the input emerge as distinct in the cluster analysis, and second, differences between tune pairs that cluster together (within-cluster) are not enhanced as a function of metrical structure. This begs the following question: are there any measurable differences between these within-cluster tunes? If yes, it should be confirmed that this difference truly is not enhanced as a function of metrical structure. Given the way the models presented in this section are structured, evidence for enhancement would be detected as the presence of an interaction between the tune variable and the metrical structure variable, showing that the difference between tunes becomes larger across the ordered metrical structures.

3.3.1 Ending F0

The first analysis in this section compares the ending F0 values of three pairs of tunes that group together in the cluster analysis in each experiment, in order to test for possible enhancement effects in this parameter. The three pairs compared for the monotonal accent experiment are {H*HH, H*HL}, {H*LH, H*LL}, and {L*HL, L*LH}, each of which has previously been shown to vary in ending F0 (e.g., Cole et al. 2023 and see **Figure 2**). As described in Section 2.4.4, we ran separate models for each tune pair. **Figure 9A** shows the values grouped by pair, with the metrical structure on the x axis. Note that, visually, enhancement would appear as a larger separation between the ending F0 values of a given pair of tunes, from left to right across the metrical conditions on the x-axis.

The model examining ending F0 in {H*HH, H*HL}, finds a credible difference between tunes, with the marginal effect showing that overall H*HH has higher ending F0 ($\hat{\beta} = 0.49$, 95CrI = [0.36,0.62], pd = 100), in line with Cole et al. (2023). There was no main effect of metrical structure ($\hat{\beta} = 0.02$, 95CrI = [–0.02,0.05], pd = 85), or interaction between metrical structure and tune ($\hat{\beta} = -0.02$, 95CrI = [–0.08,0.04], pd = 74). The estimated marginal effect of tune within each metrical condition confirms that there was a credible difference in each (all pds = 100), consistent with the overall effect.

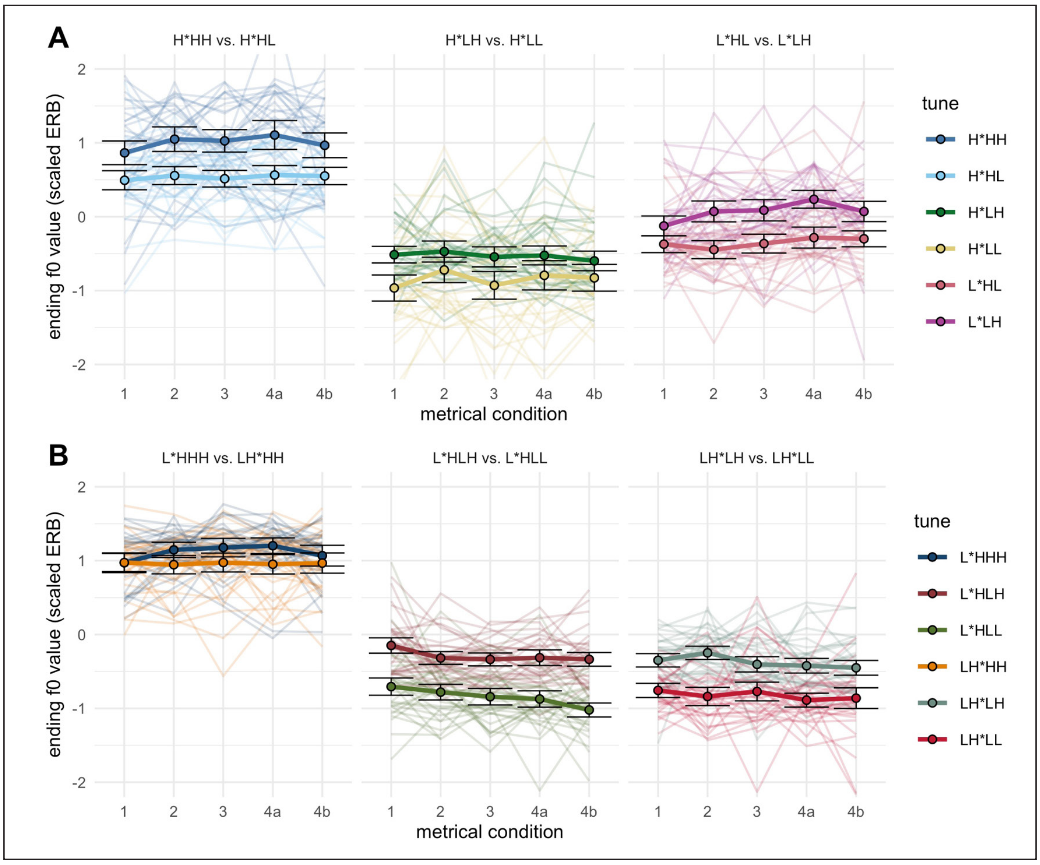


Figure 9: Scaled ERB for pairs of tunes compared in the ending F0 analysis, for the monotonal accent experiment (Panel A), and the bitonal accent experiment (Panel B), across metrical condition (x axis). The tunes in each pair are labeled at the top of each Panel and indicated by line coloration with the legend at right. Points and thicker lines are the means by tune, with error bars showing 95% CI computed from the empirical data. Lighter lines show individual participant means.

The models examining the two additional tune pairs find essentially the same result. The model for {H*LH, H*LL} finds that H*LH has higher ending F0 (marginal effect $\hat{\beta}$ = 0.29, 95CrI = [0.16,0.42], pd = 99). There was no main effect of metrical structure ($\hat{\beta}$ = –0.00, 95CrI = [–0.04,0.03], pd = 57), or interaction between metrical structure and tune ($\hat{\beta}$ = 0.03, 95CrI = [–0.03,0.10], pd = 85). The estimated marginal effect of tune within each metrical condition confirms that there was a credible difference in each (all pds > 98). The model for {L*HL, L*LH} finds that L*LH has higher ending F0 (marginal effect $\hat{\beta}$ = –0.40, 95CrI = [–0.55,–0.27], pd = 100). In a departure from the other two pairs, there was a credible main effect of metrical structure, a small positive effect ($\hat{\beta}$ = 0.04, 95CrI = [0.01,0.07], pd = 99), indicating that ending F0 overall gets higher in both tunes across metrical structures. Importantly however, there was no interaction with metrical structure and tune ($\hat{\beta}$ = 0.03, 95CrI = [–0.03,0.10], pd = 87). The estimated marginal effect of tune within each metrical condition confirms that there was a credible difference in each (all pds = 100). The analyses for the monotonal accent experiment thus lead to two conclusions: For each within-cluster pair there is indeed a difference in ending F0. Secondly, this difference is *not enhanced* as a function of metrical structure, as shown by the lack of an interaction for each pair of tunes.

The data for the bitonal accent experiment, shown in **Figure 9B**, reveals an analogous pattern. For completeness we also consider the two tunes in the bitonal accent experiment that clustered together in the rising cluster B, and differed only in pitch accent: {L*HHH, LH*HH}. These two tunes had the same ending F0 value in the stimuli, and accordingly, we do not expect them to differ in ending F0. However, the marginal effect for tune in the model finds that there is indeed a small difference whereby L*HHH has higher ending F0 than LH*HH ($\hat{\beta}$ = –0.16, 95CrI = [–0.28,–0.06], pd = 100). However, there was no main effect of metrical structure ($\hat{\beta}$ = 0.02, 95CrI = [–0.02,0.05], pd = 83), nor an interaction with tune ($\hat{\beta}$ = 0.03, 95CrI = [–0.03,0.09], pd = 80). The estimated marginal effect of tune within each metrical condition confirms that there was a credible difference in each except for the 1 syllable metrical condition (all other pds > 98).

The other tune comparisons are motivated by the difference between LH and LL edge tones, compared for each pitch accent in the bitonal accents experiment. The model examining ending F0 in {L*HLH, L*HLL} finds that L*HLH has higher F0 (marginal effect $\hat{\beta}$ = 0.49, 95CrI = [0.38,0.62], pd = 100). The model also showed a main effect of metrical structure: a small negative effect which suggests that ending F0 (for both tunes) is lower as a function of increasing metrical structure ($\hat{\beta}$ = –0.05, 95CrI = [–0.08,–0.03], pd = 100). However there was no interaction between metrical structure and tune ($\hat{\beta}$ = –0.03, 95CrI = [–0.07, 0.01], pd = 92). The estimated marginal effect of tune within each metrical condition confirms that there was a credible difference in each (all pds = 100). Finally, the model examining ending F0 in {LH*LH, LH*LL} finds that LH*LH has higher ending F0 (marginal effect $\hat{\beta}$ = 0.40, 95CrI = [0.27,0.51], pd = 100). There was some weaker evidence for a main effect of metrical structure, which showed a small negative effect ($\hat{\beta}$ = –0.03, 95CrI = [–0.05,0.00], pd = 97). As with all previous models however, there was no interaction between tune and metrical structure ($\hat{\beta}$ = 0.01, 95CrI = [–0.04,0.06], pd = 69). The estimated marginal effect of tune within each metrical condition confirms that there was a credible difference in each (all pds = 100).

The ending F0 analyses from both experiments thus allow us to conclude that tunes that group in the same cluster (the within-cluster tunes) do indeed show measurable differences in ending F0, but crucially, this difference is not enhanced as a function of metrical structure for any tune pair.

3.3.2 TCoG

The TCoG analysis examined tune pairs of interest, again informed by the clustering results. As with the preceding ending F0 results, we were interested to test for possible enhancement effects for within-cluster tune pairs with the goal of confirming a true lack of these effects as suggested by the RMSD and ending F0 analyses. Recall that we limited our comparisons in two ways: first, we did not consider the 1 or 4b metrical condition as discussed in Section 2.4.4. We also did not consider tunes that are predicted to have more than one F0 mass, namely H*LH in the monotonal accent experiment and LH*LH and L*HLH in the bitonal accent experiment.

For the monotonal accent experiment, we were interested in the difference between two within-cluster pairs of tunes. {H*HH, H*HL} was the first pair, which has previously been shown to differ the shape of the rise to the F0 maximum, with H*HH showing a scooped rise (= later TCoG) and H*HL showing a more domed rise (Cole et al. 2023). As shown in the left portion of **Figure 10**, Panel A, this was reflected in a later TCoG for H*HH as compared to H*HL (marginal effect $\hat{\beta}$ = 21, 95CrI = [12,28], pd = 100). Marginal estimates for tune within each metrical condition were also credible (pds = 100), and there was a main effect of metrical structure showing later TCoG across the ordered metrical structure conditions ($\hat{\beta}$ = 61, 95CrI = [55,67], pd = 100). There was no interaction between tune and metrical structure ($\hat{\beta}$ = 1, 95CrI = [–9,10], pd = 55), showing a lack of enhancement of TCoG differences, which can be seen visually in that tunes are not more differentiated as metrical structure increases. The second pair of tunes that we compared from the monotonal accent experiment was {L*HL, L*LH}, which our previous work shows as differing in the alignment of the F0 valley of the tune (Cole et al. 2023), leading us to predict that TCoG should be later for the later-aligned valley of L*LH. This prediction was confirmed, as shown in the right portion of **Figure 10A**, where L*LH has later TCoG as compared to L*HL (marginal effect $\hat{\beta}$ = –24, 95CrI = [–36,–11], pd = 100). Marginal estimates for tune within each metrical condition were also each credible (pds = 100), and there was a main effect of metrical structure, showing that increasing metrical structure leads to later TCoG ($\hat{\beta}$ = 70, 95CrI = [62,79], pd = 100). There was no interaction between tune and metrical structure ($\hat{\beta}$ = 2, 95CrI = [–11,14], pd = 62), again showing a lack of enhancement of TCoG differences.

The results for the two pairs of tunes from the bitonal accent experiment reveal the same pattern, shown in Panel B of **Figure 10**. We first compared TCoG as a temporal property distinguishing pitch accent alignment in {LH*LL, L*HLL}. As would be predicted, L*H showed a later alignment than LH* (marginal effect $\hat{\beta}$ = 27, 95CrI = [18,34], pd = 100). Marginal estimates for tune within each metrical condition were also each credible (pds = 100). The same pattern of later TCoG across the ordered metrical structure conditions was also apparent ($\hat{\beta}$ = 52, 95CrI = [47,57], pd = 100), again with no interaction between tune and metrical structure, therefore showing a lack of enhancement ($\hat{\beta}$ = –2, 95CrI = [–10,7], pd = 66). The case was the same for the two pitch accents in the context of a HH edge tone sequence: {LH*HH, L*HHH}. L*H showed a later TCoG overall (marginal effect $\hat{\beta}$ = 21, 95CrI = [13,30], pd = 100), and within each metrical condition (pds = 100). Increasing metrical structure once again led to later TCoG ($\hat{\beta}$ = 70, 95CrI = [64,76], pd = 100), with no interaction between tune and metrical structure ($\hat{\beta}$ = –2, 95CrI = [–10,6], pd = 66).

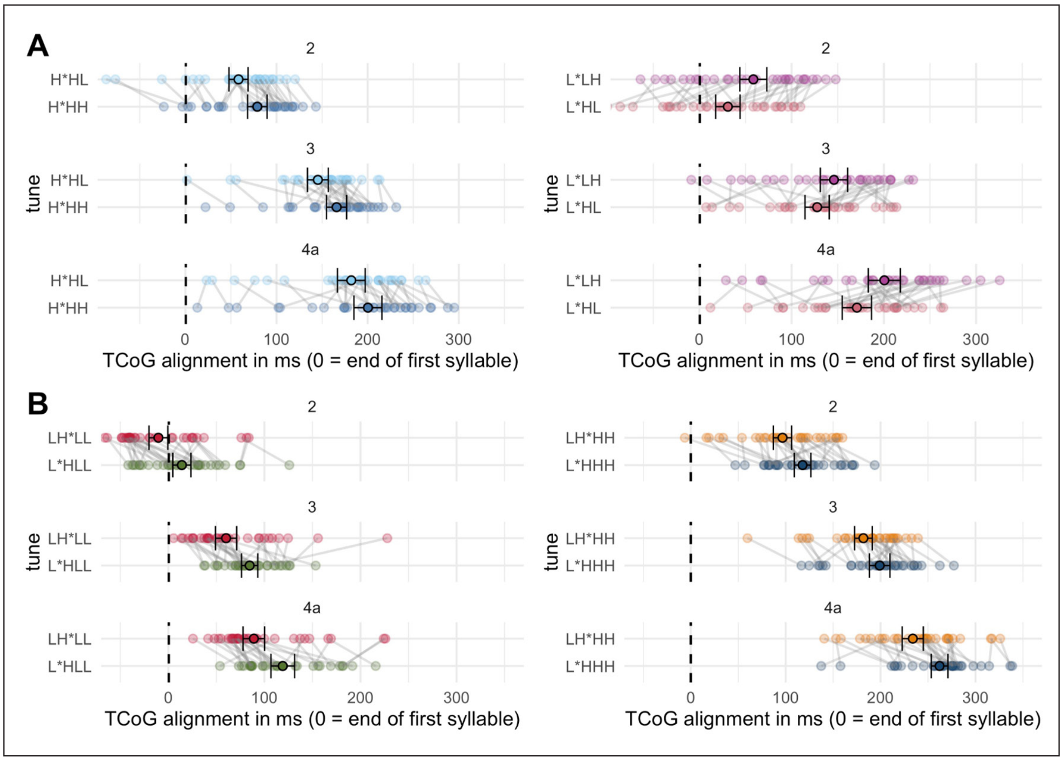


Figure 10: Temporal TCoG for two pairs of tunes in the monotonal accent experiment (Panel A), and the bitonal accent experiment (Panel B). TCoG is shown on the x axis with the relevant tune pair levels at left. The metrical condition is above each Panel (conditions 2-4a only, see text). The dashed vertical line indicates the boundary between the first and second syllable. The solid points are means and 95% CI computed from the data, the lighter points and connecting lines are individual participant means.

The TCoG analyses thus show clear distinctions in tonal timing for certain tune pairs that cluster together, but these distinctions are small, and are not enhanced as a function of metrical structure. In the monotonal accent experiment, the observed differences corresponded to the overall disposition of a higher F0 mass in a tune that rises throughout or at the end of the trajectory, where H*HH showed later TCoG than H*HL, and L*LH showed later TCoG than L*HL. However, these differences in TCoG critically did not get larger across the ordered metrical structures. The same pattern appears for the pitch-accent based distinctions in tonal timing in the bitonal accent experiment, where L*H has later TCoG than LH* in both HH and LL edge tone contexts. However, this difference also does not get larger across the ordered metrical structure.

Synthesising the two analyses in this section, the findings for TCoG are similar to those for ending F0 in that there are measurable differences between pairs of tunes that tend to cluster together. However, speakers *do not enhance* these differences as a function of metrical structure. Therefore, a lack of enhancement in same-clustering tunes, as found in the RMSD analysis, is confirmed, as is the presence of smaller and non-enhanced differences for the tune pairs examined.

4 Conclusions

The present study examined how intonational tunes differed from one another across five different metrical conditions which varied in terms of syllable count, using an imitative speech production paradigm. Our basic question was the following: do we find enhancement among the tunes, each of which are proposed categories in the AM model of Mainstream American English? We began with a clustering analysis that revealed, for each experiment, that the six tunes were not well differentiated from each other. Additionally, the clustering partition predicted whether or not tunes became *more different* from one another as a function of metrical structure. Intriguingly, tunes that clustered together did not, overall, become more separated in F0 space across metrical conditions. Follow-up analyses of ending F0 and TCoG confirmed this, showing that there were indeed small but detectable differences between tunes that were grouped in the same cluster, though those differences did not change across metrical conditions. In this sense, we have evidence that our metrical structure manipulation did successfully induce enhancement, but only for some tunes. Here we consider what these results imply both methodologically and go on to discuss theoretical implications in Section 4.1.

First, we consider a methodological perspective. What does the participants' performance on this imitative task tell us about the representations being accessed therein? The data overall showed fairly consistent productions for each tune across the five metrical conditions tested. This was clearest in the bitonal accents experiment in which the optimal number of clusters (two) did not change across metrical conditions, and the partition of tunes into clusters remained essentially the same. For the monotonal accent experiment, the change in the optimal number of clusters from two to three across metrical conditions suggests some loss of distinctiveness between {H*HH, H*HL} and {L*HL, L*LH}, however, this loss is clearly not to the extent that the intonational distinctions fully disappear as is evident from **Figure 3**, wherein the overall shape of each tune does not change dramatically across metrical conditions. We conclude that, overall, speakers were successful at transposing the tune from the model to metrically different target productions, a conclusion that is reinforced by ending F0 and TCoG analyses, which each show comparable differences in these parameters for the relevant tune pairs across metrical conditions. The ending F0 analysis in particular suggests that tunes do not undergo truncation in these data. Had there been truncation, we would expect that low-ending or falling tunes would end in a lower F0 when produced over more syllables, and likewise, high-ending or rising tunes would end in a higher F0. We do see evidence for this pattern with {L*HL, L*LH}, for example, where a main effect of metrical structure was consistent with this pattern. However, the effect was quite small, and its absence for other tune pairs suggests it is not systematic.

The general success of the imitation paradigm in eliciting transposed tunes leaves open various future questions about its application to the study of intonational representation. The relative success of transferring these nuclear contours may offer a comparison case for future studies which may make use of imitation to examine how intonation is represented by listeners and speakers. For example, pre-nuclear intonational melodies have been suggested to be "ornamental" and less important in conveying intonational meaning (cf., Baumann et al. 2017; Chodroff & Cole 2018). Testing if and how pre-nuclear tunes are produced across metrical structures may offer a lens into what sorts of intonational properties are salient, and are reproduced faithfully in various metrical contexts. Zahner-Ritter et al. (2022) is a recent study that, in a similar vein, has tested distinctions among intonational categories in German in a production study where melodies were direct imitations of heard stimuli (with a 2 second lag and 500 ms sine tone). Imitation across various metrical structures and examining methodological permutations of this sort seems like a promising program of research to investigate intonational tunes and the phonological underpinnings of the system.

4.1 Two kinds of intonational distinctions

Fundamentally, the present results point to two types of distinctions in the intonational tunes produced by the speakers in our experiments. The first type of distinction is large and salient, is emergent in clustering, and is enhanced as a function of metrical structure. The second type is smaller, not emergent in clustering, and not enhanced as a function of metrical structure. Both of these types of distinctions are encoded with phonological tone labels in the ToBI annotation scheme, which formed the basis of our stimuli.

The way ToBI labels relate to the present distinctions is not entirely straightforward. In the monotonal accents experiment, we find that the small and non-enhanced distinctions for some tune pairs correspond to differences in boundary tone labels: {H*HH, H*HL}, and {H*LH, H*LL}. For the tune pair {L*HL, L*LH} the distinction is labeled as both a phrase accent and boundary tone. The pattern in the bitonal accent experiment is different, whereby the large and enhanced distinctions are based on edge tones, and more specifically the phrase accent label. Tunes that combined a bitonal pitch accent with HH edge tones were distinct from tunes with the same pitch accent and LH or LL edge tones, though the tunes ending in LH and LL clustered together. Across experiments then, there is not a clear unifying tone label that predicts whether a particular pair of tunes will be robustly distinguished, or not. It is important to note here that the clustering partition of tunes for each experiment is dependent on the set of tunes that was tested in that experiment, and future work should consider different sets of tunes in different combinations. Nevertheless, it seems reasonable to assume that if, hypothetically, pitch accent distinctions were always the distinctions which were emergent and enhanced, this would be the case across experiments. This is clearly not the case, and instead, a better understanding of what properties are shared by emergent and enhanced distinctions comes from considering the contours more holistically. In the monotonal accent experiment, the distinctions among high-rising, rising-falling, and low-to-mid rising shapes are enhanced. In the bitonal accent experiment, the distinction between rising-falling and rising is enhanced. This distinction between monotonically rising contours and contours with other shapes was emergent in our clustering analyses and also in previous work, suggesting that it constitutes a fundamental dimension in intonational melodies, shown to be emergent in iterative imitations in Braun et al. (2006), and clustering analyses in Cole et al. (2023) and Steffman et al. (2022). Distinctions between high-rising tunes and tunes with other shapes, already quite a large difference in F0 space, becomes more distinct in longer words, whereas tunes that are more similar to one another, e.g., those that are grouped together in clustering analyses do not, as shown by the RSMD, TCoG and ending-F0 results.

What does this difference in tune distinctions imply for a theory of intonational phonology, e.g., concerning the category status of pitch accents, as outlined in the Introduction? One answer to the question would be to take enhancement as a direct diagnostic for categoricity: that is, enhanced distinctions are category-level distinctions, while non-enhanced distinctions are not. From this view, following Cole et al. (2023), the tune shapes that emerge from the clustering analyses could be described as category-level distinctions, which further evidence category-like behavior in being subject to enhancement across metrical structures. Perhaps most interestingly, the second type of distinction, which does not emerge from clustering analyses and is not enhanced, would by this metric not be a category-level distinction. How then, should such “secondary” distinctions be considered? We suggest that these distinctions may best be understood as potentially meaningful, within-category, continuous variation. In other words, they constitute variation in a parameter that is heard and reproduced by speakers, and which may convey intonational meaning, but which falls along a scalar or phonetic dimension, and does not mark a categorical distinction in the intonational system (and hence, is not enhanced).

Considered in this way, the bitonal accent experiment indicates that the distinction between L+H* and L*+H pitch accents represents within-category variation, which accords with Gussenhoven (1984), who suggested that this difference should be considered as a continuous alignment parameter, not a categorical distinction. In discussing the possibility that alignment distinctions are continuous and not categorical, Pierrehumbert & Steele (1989) state that that hypothesis is difficult to assess if continuous alignment distinctions show preferred, or typical, values. Enhancement seems to offer a useful diagnostic, and by that metric, the model of alignment distinctions between L+H* and L*+H as continuous, not categorical, is supported. The enhancement results for the bitonal accents presented here also agree with the suggestion in Ladd (2022), that there may just be one phonological high accent in the system. As shown in the TCoG analysis, there were measurable differences between the two pitch accents related to F0 peak alignment, however, these differences did not grow across expanded metrical structures, and were not sufficient to group productions of the two pitch accents in the same emergent cluster. In addition to peak alignment, a second fundamental type of distinction that was not enhanced across both experiments relates to the ending F0 value for tunes that differed primarily in this parameter (a difference manifest on the final syllable of the trisyllabic word of the model tune, though this distinction does not map onto any one specification of phrase accent and/or boundary tone, as noted above). The distinction in ending F0 between similar tune pairs is clearly not enhanced, as shown by the ending F0 analysis, and, on analogy with the pitch accent distinction, may best be viewed as variation along an F0 scaling continuum. We reiterate here again, that future examination of both of these types of distinctions should consider intonational meaning, and their role in the interpretation of intonational contours.

4.2 Some limitations and future directions

Several key features of the design of this experiment should be kept in mind, both as limitations for the present study, and in possible future directions they raise for research along these lines.

Firstly, we focused here only on the possible enhancement of F0 parameters. While F0 has certainly been a central object of study in the intonation literature for many years, and the only acoustic parameter explicitly defined by the AM model, we do not wish to imply that it is the only possible correlate of these tunes. Secondary, and less-described, cues to tune distinctions (e.g., intensity and voice quality) may also certainly be enhanced. An analogy in the segmental domain would be the enhancement of vowel duration as a cue to voicing distinctions and vowel category in under prosodic prominence in American English (De Jong 2004). This represents an important avenue for extending the present results, though further fundamental research into non-F0 correlates of these tunes strikes us as a needed preliminary in this vein. As it pertains to duration, we can note however, that durational distinctions between tunes were not large, nor did they become larger with increased metrical structure, as shown in **Figures 3** and **4**.

Secondly, we examined enhancement in just one light: as a function of changing syllable count and metrical structure. This is just one possible way of testing enhancement, and future work should examine how these tunes, and others, are produced in different contexts which may be enhancing. The production of tunes in different emotional contexts, or in hyperarticulated speech (cf. Arvaniti & Garding 2007, and Schertz 2013 for vowel contrasts) strikes us as a promising test case. Observing if the same distinctions we examine here are enhanced under these, very different, contextual manipulations, would provide deeper insight into how enhancement operates. In this vein, because the stimuli in these experiments were presented in the absence of any discourse context, this is fundamentally a study of intonation form, as perceived and produced by speakers. To gain a fuller understanding, the pragmatic and discourse functions of each of these tunes must be considered. In addition to metrical enhancement of the sort studied here, it may be possible that discourse contexts that are appropriate for a given tune could also serve an enhancing role, making tunes more distinct from one another. We suggest that this will be an important next step in furthering the present results.

Another consideration which we believe is a promising extension of these findings is to consider dialectal variation. As we tested only the categories proposed for "mainstream" American English and we recruited speakers from multiple regions in the US, we are not in a position to make claims about the dialect-specificity of these results. The realization of intonation elements (especially pitch accents) has been shown to differ across regional dialects and varieties of American English (Arvaniti & Garding 2007; Burdin et al. 2022). Examining how the effects shown here play out across varieties is thus another important extension of the present results. The pattern of results we show here, whereby only emergent and large distinctions are enhanced, strikes us as an intriguing test for intonation systems across dialects: are the same distinctions emergent? Are these and only these distinctions enhanced? Pursuing this link between emergent distinctions and their enhancement across dialects will allow for a more holistic picture of intonational categories. This approach, however, would necessitate a consensus understanding of the intonational elements and proposed categories in particular variety, which may or may not be commensurate with MAE labels (cf. Burdin et al. 2022). In this same vein, considering how individual speakers partition, and enhance, distinctions within the space of tunes may further be insightful for the study of variation in and individual differences in intonation (cf. Niebuhr et al. 2011; Cole & Steffman 2021).

4.3 Conclusions

In summary, the present results, in line with other recent work (Steffman et al. 2022; 2024; Cole et al. 2023), show that not all of the predicted tune contrasts from the AM model are emergent in imitated productions, based on clustering analyses. From two experiments which together tested distinctions among 12 tunes, our results suggest three emergent tunes, described in terms of their shape as high-rising, rising-falling, low-rising, with additional fine-grained distinctions in rise shape (measured here by TCoG) and ending F0 within these emergent tunes. The most central finding, present in both experiments, is that only those distinctions that emerge from clustering analyses are enhanced. Other tune distinctions predicted by the AM model fail to emerge in clustering analyses of the data, and are instead reflected in F0 distinctions of small magnitude, which do not vary as a function of metrical structure. This distinction between emergent and non-emergent tunes was interpreted as a fundamental dissociation: if enhancement targets phonological categories, smaller F0 differences between imitated productions may best be understood as encoding (potentially meaningful) within-category phonetic variation. Considering the intonational distinctions in American English along these lines leads to a possible reconciliation among divergent claims about the nature and number of intonational distinctions in the language (e.g., Gussenhoven 1984; Calhoun 2012; Ladd 2022; Steffman et al. 2024). Future work testing this hypothesis will build our understanding of intonational phonology and how intonation is perceived and produced by language users.

At a more general level, we have shown that testing enhancement in the manner we did can offer insights into the study of intonational phonology and the categories in a system. We believe fruitful extensions of this general approach could take the same line of inquiry to other intonational systems, or more targeted examinations of intonational elements (e.g., pitch accents, or boundary tones only). Beyond intonational phonology, the F0 analysis protocols here may be considered a way to study lexical tone contrasts, and the possible enhancement of lexical tone distinctions across various contexts. Most broadly we believe that testing enhancement, in combination with the examination of emergent or “bottom-up” distinctions, is a powerful tool that has the potential to provide deep insights into phonological systems. This may be especially true in cases where analyses diverge, where the nature of the system is not well understood, and for the development of phonological analyses and descriptions of un(der)-described languages.

Acknowledgments

We are thankful to Chun Chan for technical help, and to the Northwestern Prosody and Speech Dynamics lab for helpful discussion. This project was supported by NSF BCS-1944773.

Competing interests

The authors have no competing interests to declare.

References

Arvaniti, Amalia & Garding, Gina. 2007. Dialectal variation in the rising accents of American English. *Papers in Laboratory Phonology* 9. 547–576.

Barnes, Jonathan & Brugos, Alejna & Veilleux, Nanette & Shattuck-Hufnagel, Stefanie. 2021. On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics* 85. 101020. DOI: https://doi.org/10.1016/j.wocn.2020.101020

Barnes, Jonathan & Veilleux, Nanette & Brugos, Alejna & Shattuck-Hufnagel, Stefanie. 2012. Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology* 3(2). 337–383. DOI: https://doi.org/10.1515/lp-2012-0017

Baumann, Stefan & Mertens, Jane & Kalbertodt, Janina & Phonetik, IfL. 2017. How'ornamental'are German prenuclear accents. *Oral presentation at PaM17* 4.

Beckman, Jill & Helgason, Pétur & McMurray, Bob & Ringen, Catherine. 2011. Rate effects on Swedish VOT: Evidence for phonological overspecification. *Journal of Phonetics* 39(1). 39–49. DOI: https://doi.org/10.1016/j.wocn.2010.11.001

Beckman, Mary E. & Pierrehumbert, Janet B. 1986. Intonational structure in Japanese and English. *Phonology* 3. 255–309. DOI: https://doi.org/10.1017/S095267570000066X

Boersma, Paul & Weenink, David. 2020. Praat: doing phonetics by computer (version 6.1.09). http://www.praat.org.

Bolinger, Dwight. 1961. *Generality, gradience, and the all-or-none*. Mouton.

Braun, Bettina & Kochanski, Greg & Grabe, Esther & Rosner, Burton S. 2006. Evidence for attractors in English intonation. *The Journal of the Acoustical Society of America* 119(6). 4006–4015. DOI: https://doi.org/10.1121/1.2195267

Burdin, Rachel Steindel & Holliday, Nicole R. & Reed, Paul E. 2022. American English pitch accents in variation: Pushing the boundaries of mainstream American English-ToBI conventions. *Journal of Phonetics* 94. 101163. DOI: https://doi.org/10.1016/j.wocn.2022.101163

Bürkner, Paul-Christian. 2017. brms: An r package for Bayesian multilevel models using stan. *Journal of Statistical Software* 80. 1–28. DOI: https://doi.org/10.18637/jss.v080.i01

Bürkner, Paul-Christian & Charpentier, Emmanuel. 2020. Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology* 73(3). 420–451. DOI: https://doi.org/10.1111/bmsp.12195

Calhoun, Sasha. 2012. The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics* 40(2). 329–349. DOI: https://doi.org/10.1016/j.wocn.2011.12.001

Caliński, Tadeusz & Harabasz, Jerzy. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3(1). 1–27. DOI: https://doi.org/10.1080/03610927408827101

Cho, Taehong. 2004. Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics* 32(2). 141–176. DOI: https://doi.org/10.1016/S0095-4470(03)00043-3

Cho, Taehong. 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of/ɑ, i/in English. *The Journal of the Acoustical Society of America* 117(6). 3867–3878. DOI: https://doi.org/10.1121/1.1861893

Cho, Taehong & Kim, Daejin & Kim, Sahyang. 2017. Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics* 64. 71–89. DOI: https://doi.org/10.1016/j.wocn.2016.12.003

Chodroff, Eleanor Rosalie & Cole, Jennifer. 2018. Information structure, affect, and prenuclear prominence in American English. In *Proceedings of INTERSPEECH*. 1848–1852. International Speech Communication Association. DOI: https://doi.org/10.21437/Interspeech.2018-1529

Chodroff, Eleanor Rosalie & Cole, Jennifer. 2019. 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, Jennifer & Kim, Heejin & Choi, Hansook & Hasegawa-Johnson, Mark. 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics* 35(2). 180–209. DOI: https://doi.org/10.1016/j.wocn.2006.03.004

Cole, Jennifer & Steffman, Jeremy. 2021. The primacy of the rising/non-rising dichotomy in American English intonational tunes. In *Proceedings of the 1st International Conference on Tone and Intonation*, 122–126. DOI: https://doi.org/10.21437/TAI.2021-25

Cole, Jennifer & Steffman, Jeremy & Shattuck-Hufnagel, Stefanie & Tilsen, Sam. 2023. Hierarchical distinctions in the production and perception of nuclear tunes in American English. *Laboratory Phonology* 14(1). DOI: https://doi.org/10.16995/labphon.9437

De Jong, Kenneth. 2004. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics* 32(4). 493–516. DOI: https://doi.org/10.1016/j.wocn.2004.05.002

Dilley, Laura C. 2005. *The phonetics and phonology of tonal systems*: Massachusetts Institute of Technology dissertation.

Dilley, Laura C. & Heffner, Christopher C. 2013. The role of f0 alignment in distinguishing intonation categories: evidence from American English. *Journal of Speech Sciences* 3(1). 3–67. DOI: https://doi.org/10.20396/joss.v3i1.15039

D'Imperio, Mariapaola & German, James. 2015. Phonetic detail and the role of exposure in dialect imitation. In *18th International Congress of Phonetic Sciences*.

Garellek, Marc. 2014. Voice quality strengthening and glottalization. *Journal of Phonetics* 45. 106–113. DOI: https://doi.org/10.1016/j.wocn.2014.04.001

Genolini, Christophe & Falissard, Bruno. 2011. KmL: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine* 104(3). e112–e121. DOI: https://doi.org/10.1016/j.cmpb.2011.05.008

Grabe, Esther. 1998. Pitch accent realization in English and German. *Journal of Phonetics* 26(2). 129–143. DOI: https://doi.org/10.1006/jpho.1997.0072

Grabe, Esther & Post, Brechtje & Nolan, Francis & Farrar, Kimberley. 2000. Pitch accent realization in four varieties of British English. *Journal of Phonetics* 28(2). 161–185. DOI: https://doi.org/10.1006/jpho.2000.0111

Grice, Martine. 2017. *The intonation of interrogation in Palermo Italian*. De Gruyter Mouton.

Gussenhoven, Carlos. 1984. *On the grammar and semantics of sentence accents*. De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110859263

Gussenhoven, Carlos. 2006. Experimental approaches to establishing discreteness of intonational contrasts. *Methods in Empirical Prosody Research,* 321–334. DOI: https://doi.org/10.5070/P71FP2N3QQ

Im, Suyeon & Cole, Jennifer & Baumann, Stefan. 2023. Standing out in context: Prominence in the production and perception of public speech. *Laboratory Phonology* 14(1). DOI: https://doi.org/10.16995/labphon.6417

Kawahara, Hideki & Cheveigné, Alain de & Banno, Hideki & Takahashi, Toru & Irino, Toshio. 2005. Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Ninth european conference on speech communication and technology*. DOI: https://doi.org/10.21437/Interspeech.2005-335

Keyser, Samuel Jay & Stevens, Kenneth Noble. 2006. Enhancement and overlap in the speech chain. *Language,* 33–63. DOI: https://doi.org/10.1353/lan.2006.0051

Kim, Sahyang & Kim, Jiseung & Cho, Taehong. 2018. Prosodic-structural modulation of stop voicing contrast along the vot continuum in trochaic and iambic words in American English. *Journal of Phonetics* 71. 65–80. DOI: https://doi.org/10.1016/j.wocn.2018.07.004

Ladd, D. R. 2022. The trouble with ToBI. In Barnes, Jonathan & Shattuck-Hufnagel, Stefanie (eds.), *Prosodic theory and practice*, 247–258. MIT Press. DOI: https://doi.org/10.7551/mitpress/10413.003.0009

Ladd, D. R. & Schepman, Astrid. 2003. "sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of Phonetics* 31(1). 81–112. DOI: https://doi.org/10.1016/S0095-4470(02)00073-6

Lenth, Russell V. 2021. *emmeans: Estimated marginal means, aka least-squares means*. https://CRAN.R-project.org/package=emmeans. R package version 1.7.1-1.

McAuliffe, Michael & Socolof, Michaela & Mihuc, Sarah & Wagner, Michael & Sonderegger, Morgan. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of INTERSPEECH*, 498–502. DOI: https://doi.org/10.21437/Interspeech.2017-1386

Moulines, Eric & Charpentier, Francis. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9(5–6). 453–467. DOI: https://doi.org/10.1016/0167-6393(90)90021-Z

Niebuhr, Oliver & d'Imperio, Mariapaola & Fivela, Barbara Gili & Cangemi, Francesco. 2011. Are there "shapers" and "aligners"? Individual differences in signalling pitch accent category. In *Proceedings of the 17th International Congress of Phonetic Sciences*, 120–123.

Pierrehumbert, Janet B. 1980. *The phonology and phonetics of English intonation*: Massachusetts Institute of Technology dissertation.

Pierrehumbert, Janet B. & Hirschberg, Julia B. 1990. The meaning of intonational contours in the interpretation of discourse. In Cohen, Philip R. & Morgan, Jerry & Pollack, Martha E. (eds.), *Intentions in Communication*, MIT Press. DOI: https://doi.org/10.7551/mitpress/3839.003.0016

Pierrehumbert, Janet B. & Steele, Shirley A. 1989. Categories of tonal alignment in English. *Phonetica* 46(4). 181–196. DOI: https://doi.org/10.1159/000261842

Posit team. 2023. *Rstudio: Integrated development environment for r*. Posit Software, PBC Boston, MA. http://www.posit.co/.

R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Rathcke, Tamara. 2013. On the neutralizing status of truncation in intonation: A perception study of boundary tones in German and Russian. *Journal of Phonetics* 41(3–4). 172–185. DOI: https://doi.org/10.1016/j.wocn.2013.01.003

Sadeghi, Vahid. 2023. Phonetic effects of tonal crowding in persian polar questions. *Language and Speech* 00238309231213580. DOI: https://doi.org/10.1177/00238309231213580

Schertz, Jessamyn. 2013. Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics* 41(3–4). 249–263. DOI: https://doi.org/10.1016/j.wocn.2013.03.007

Shue, Yen-Liang & Keating, Patricia & Vicenik, Chad & Yu, Kristine. 2009. Voicesauce. *p. Program available online at http://www. seas. ucla. edu/spapl/voicesauce/. UCLA* .

Steffman, Jeremy & Cole, Jennifer. 2022. An automated method for detecting F0 measurement jumps based on sample-to-sample differences. *JASA Express Letters* 2(11). 115201. DOI: https://doi.org/10.1121/10.0015045

Steffman, Jeremy & Cole, Jennifer & Shattuck-Hufnagel, Stefanie. 2024. Intonational categories and continua in American English rising nuclear tunes. *Journal of Phonetics* 104. 101310. DOI: https://doi.org/10.1016/j.wocn.2024.101310

Steffman, Jeremy & Shattuck-Hufnagel, Stefanie & Cole, Jennifer. 2022. The rise and fall of American English pitch accents: Evidence from an imitation study of rising nuclear tunes. *Proceedings of Speech Prosody,* 857–861. DOI: https://doi.org/10.21437/SpeechProsody.2022-174

Stevens, Kenneth N. & Keyser, Samuel Jay. 1989. Primary features and their enhancement in consonants. *Language,* 81–106. DOI: https://doi.org/10.2307/414843

Sundberg, Johan. 1973. Data on maximum speed of pitch changes. *Speech transmission laboratory quarterly progress and status report* 4. 39–47.

Veilleux, Nanette & Shattuck-Hufnagel, Stefanie & Brugos, Alejna. 2006. Transcribing prosodic structure of spoken utterances with ToBI (Version 6.911). Massachusetts Institute of Technology: MIT OpenCourseWare. https://ocw.mit.edu.

Yu, Jenny & Zahner, Katharina. 2018. Truncation and compression in Southern German and Australian English. In *Proceedings of INTERSPEECH*, 1833–1837. DOI: https://doi.org/10.21437/Interspeech.2018-2513

Zahner-Ritter, Katharina & Einfeldt, Marieke & Wochner, Daniela & James, Angela & Dehé, Nicole & Braun, Bettina. 2022. Three kinds of rising-falling contours in German wh-questions: Evidence from form and function. *Frontiers in Communication* 58. DOI: https://doi.org/10.3389/fcomm.2022.838955