VARIABILITY, OVERLAP, AND CUE TRADING IN INTONATION

Amalia Arvaniti[1], Argyro Katsika[2], Na Hu[1]

[1]*CLS, Radboud University;* [2]*Department of Linguistics, UC Santa Barbara*

amalia.arvaniti@ru.nl; argyro@ucsb.edu; na.hu@ru.nl

ABSTRACT

We modelled the Greek H*, L+H*, and H*+L pitch accents using Functional Principal Component Analysis followed by statistical modelling and curve reconstruction. The accents were distinguished by F0 height and shape. The data also exhibited cue trading between F0 and duration, as well as systematic context-driven variation and general variability which led to category overlap comparable to that reported for vowel contrasts. These findings indicate that intonation categories are more similar to segmental categories than previously thought, supporting the view that the study of intonation phonetics and phonology should follow the same principles as the study of segments. [1]

Keywords: intonation, variability, variation, F0, Functional Principal Component Analysis, cue trading, category overlap

[1] We thank our participants who generously donated their time to this project, and Katherine Marcoux, Georg Lohfink, and Yang Li who handled preliminary data processing and analysis. We also thank two anonymous reviewers, and Andries Coetzee, Shelome Gooden, and Anne-Michelle Tessier, for their helpful comments and suggestions. The financial support of the European Science Foundation through grant #ERC-ADG-835263 to Amalia Arvaniti, and intramural support from the University of Kent to Amalia Arvaniti in the early stages of this work, is hereby gratefully acknowledged.

1 INTRODUCTION

This paper deals with the phonological representation and phonetic realization of intonation. As the term INTONATION is used to refer to several entities and concepts, we start with a definition. Intonation refers to those language-specific and systematic modulations of F0 that have the utterance as their domain and fulfil several grammatical functions, namely encoding pragmatic information and focus and marking phrasal boundaries (Gussenhoven 2004, Hirschberg 2004, Ladd 2008, Beckman & Venditti 2010, Krivokapić 2022, Arvaniti 2022a). Thus, the present paper does not explicitly address lexical uses of F0, such as tone and lexical pitch accent, though the parallels between intonation and these lexical phenomena make our methodology suitable for the latter as well, a point we briefly discuss in section 4. In addition, intonation is not equated here with F0 or pitch, but refers only to their specific uses defined above. Indeed, as shown below, the conflation of intonation with F0 (and pitch) is one of the drawbacks of several intonation models. It has led to intonation long being seen as peripheral to linguistic inquiry and fundamentally different from segmentals. For instance, in his *Language* review of Halliday's *Intonation and grammar in British English* (Halliday 1967), Crystal notes that '[i]t certainly cannot be taken for granted that intonation is systematic in any a-priori (grammatical) sense: but this seems to be what Halliday is doing' (Crystal 1969: 385). In 1978, Bolinger described intonation as a 'half-tamed savage' (Bolinger 1978: 475), a questionable metaphor that became popular when adopted by Gussenhoven (2004, chap. 4), even though this author took pains to distinguish 'structural intonation' from 'purposeful variation in phonetic implementation' (Gussenhoven 2004: 58).

In contrast to the widespread positioning of intonation at the linguistic periphery, it is now increasingly recognized that intonation plays a pivotal role in speech production , processing, and acquisition and is essential for communication (among many, Carbary et al. 2015 and Hellbernd & Sammler 2018 on speech processing by neurotypical adults, Braun & Tagliapietra 2011, and Schmidt et al. 2020 on intonation in L2, Kottmann, Wanner & Wermke 2023 on melody in infant vocalizations, and Grice et al. 2023 on intonation in Autism Spectrum Disorder). Therefore, a better understanding of how intonation is phonologically structured and phonetically realized is imperative but still far from achieved.

A critical yet unresolved issue relates to the need to both account for phonetic variability and reach phonological abstractions useful for capturing generalizations about intonation form and meaning. This problem is not new or unique to intonation, as it applies to all levels of speech analysis (see e.g. Ladd 2014, Bürki 2018). It has been, however, particularly challenging in the study of intonation for several reasons.

First, because F0 curves do not present obvious discontinuities (other than those due to voicelessness), one cannot carve them up by simply following some standard segmentation criteria similar to those used for vowels and consonants (e.g. Machač & Skarnitzl 2009). The difficulty is illustrated in Figure 1, which shows the spectrograms of two utterances and their corresponding F0 contours. The spectrograms show clear pattern changes such that an untrained individual may not be able to determine which sounds are produced but they can determine where changes take place. The same does not apply to the F0 contours, however.

Further, F0 exhibits variability due to a number of sources. One crucial source is linguistic context because it leads to systematic changes to F0 contours which are, nevertheless, treated by speakers and listeners as instances of the same tune. This is the case with the tunes in Figure 1, which are both instances of the most frequent English tune used for broad focus, that is, with statements providing new information. Though the pragmatics of the contours is the same, their phonetic form presents marked differences: in *a yew* (capital EW in Figure 1a), the stressed vowel is rising-falling, while in *an anemone* (capital E in Figure 1a) it is rising; in *a yew*, the fall ends practically at the end of the utterance, while in *an anemone*, the fall ends earlier so the second half of the last vowel is flat.[2] These differences – discussed in detail in section 1.2 – are examples of changes dependent on linguistic content and collectively described by Arvaniti and Ladd (2009) as lawful variability. Here we follow Bürki (2018) and use the term VARIATION instead, to distinguish linguistically driven changes from general VARIABILITY. As mentioned, despite their differences, speakers recognize the contours in Figure 1 as instances of the same tune. A theory of intonation must be able to capture this critical relationship between how tunes are realized and how they are categorized by speakers. As argued in section 1.2, this can only be achieved if tunes are broken down into component events.

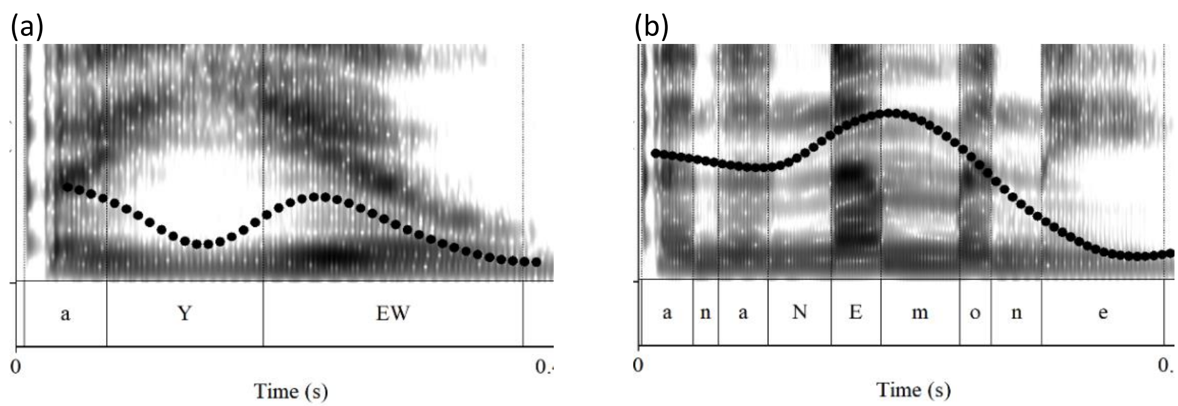


FIGURE 1. Spectrograms and F0 contours of (a) ‘a yew’ and (b) ‘an anemone’ produced by the same female speaker as responses to ‘what’s this plant?’.

Linguistically-driven variation can be juxtaposed to noise. Noise is inevitable, since no two utterances are phonetically identical, and interspeaker variation can be substantial (Niebuhr et al. 2011, Grice et al. 2017). Figure 2a illustrates noise with time-warped F0 contours of five repetitions of the phrase in Figure 1a, all produced by the same speaker during the same recording session: the overall shape is consistent but there are differences in the exact location and scaling of F0 minima and maxima.

[2] The F0 dip at the onset of *yew* is due to the glide [j] suppressing F0. Such segment-driven F0 perturbations are known as MICROPROSODY and are ignored by listeners during the processing of intonation (’t Hart et al. 1990, chap. 2; Ladd 2008, chap. 1). For this reason, such effects are not discussed here.

The source of intonation variability that is least understood and most vexing is GRADIENCE. Gradience can reflect dialectal differences (Atterer & Ladd 2004 on German, Holliday 2019 on African American English), or be the outcome of style, affect, or politeness (Nadeu & Prieto 2011 on Catalan, Henriksen 2013 on Spanish, Gryllia, Baltazani & Arvaniti 2018 on Greek). Thus, gradience may be meaningful and important for communication – 'purposeful' as Gussenhoven (2004: 58) puts it. However, gradience does not always reflect phonological distinctions, such as a change from one accent type to another, and is not always meaningful; for example, the scaling difference between Figure 1a and 1b is most likely incidental. Thus, gradience can be difficult to disentangle from category differences, on the one hand, and noise on the other, and is possibly the main reason why intonation has long been seen as intrinsically different from segments.

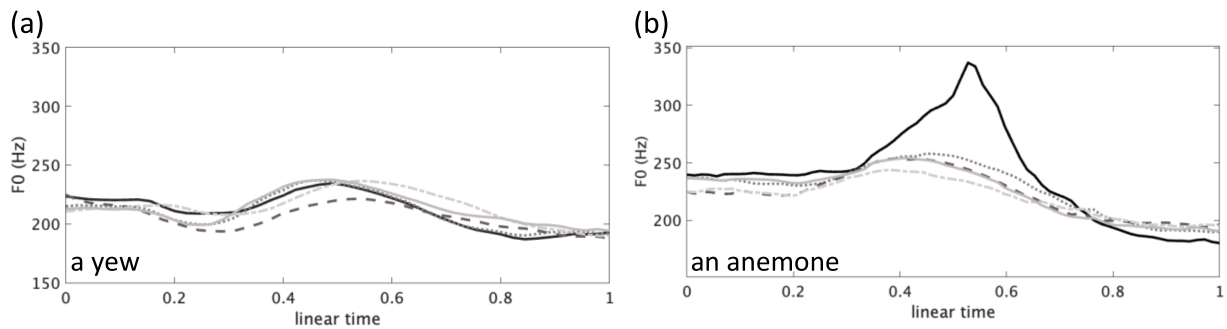


FIGURE 2. Five time-warped renditions of the tune depicted in Figure 1 produced by the same female speaker during a single recording session; time-warping performed using a non-linear time-warping procedure developed by Mark Tiede (Haskins Laboratories), following Lucero et al. (1997).

Gradience is illustrated in Figure 2b, where the peak of the solid black contour is both higher and later than in the other repetitions of *an anemone*, though they were all produced by the same speaker as responses to *what's this plant?* and thus had the same pragmatic intent. How to handle such differences remains a point of disagreement. In accounts based in the autosegmental metrical theory of intonational phonology (henceforth AM; Pierrehumbert 1980, Ladd 2008), they are likely to be treated as within-category gradience. In most British School analyses, on the other hand, the black contour would be taken to belong to a distinct category, a HIGH FALL, while the low-peak contours would be treated as LOW FALLS (O'Connor & Arnold 1973, Cruttenden 1997). Both AM and the British School approaches recognize that the difference is one of attitudinal nuance, the high fall being more lively or involved than the low fall, but while the former see this as a linguistic distinction, many AM analyses consider it PARALINGUISTIC: 'clearly meaningful but not […] organized along linguistic lines' (Ladd 2008: 34; see also Gussenhoven 2004, chap. 5). We return to this point in section 4 but stress that, independently of the answer to this particular conundrum, a desideratum for a theory of intonation must be its ability to handle gradience. A better handle on gradience would make it possible to assign the contours of Figure 2b into one or two distinct phonological categories with some certainty, and treat their variability in a principled manner.

In order to reach all of the above desiderata, it is important for any intonation theory to address the following questions.

A. How do we capture the commonalities that lead speakers of most English varieties to interpret the contours in Figures 1 and 2 as instances of the same tune?

B. How do speakers know when to use this tune, and how do they learn how to implement it over utterances that differ in length and metrical structure?

C. As linguists, how do we distinguish between variation and (inevitable) random noise?

D. Since many types of variability involve gradience, how can we make a principled distinction between categorical and gradient differences? For instance, how can we tell whether the contours depicted in Figure 2b are instances of the same tune and not of two distinct tunes?

Most models of intonation do not provide a coherent way of tackling all these questions. Before presenting our own data and analysis, we briefly review a number of established models that have played a critical role in the development of our current understanding of intonation and discuss how they handle variability and abstraction.

1.1 Treatments of phonetic variability and abstraction in intonation

The challenge of determining melodic structure while handling intonation variability has been met with a number of approaches that fall largely into two groups. The first includes models that use iconic notations to represent idealized versions of pitch contours. Figure 3 shows several such representations of the contour in Figure 1b: (a) the 'tadpole' notation of the British School, following O'Connor & Arnold (1973); (b) the IPO system first developed for Dutch ('t Hart, Collier & Cohen 1990); (c) the pitch levels of the American structuralists, following Pike (1945); (d) Bolinger's use of typesetting to imitate pitch movements (e.g. Bolinger 1986); INTSINT, in which arrows indicate the direction and height of pitch changes (e.g. Hirst & Di Cristo 1998). To a significant extent, these are all representations of F0 rather than intonation, while their focus on idealization renders variability a moot point.[3]

The goal of the second set of models is to faithfully reproduce, rather than idealize, F0 contours. For instance, Fujisaki and his colleagues model F0 contours as combinations of declining phrase commands and accent commands that disrupt the declination line (e.g. Fujisaki 1983, Gu, Hirose & Fujisaki 2007). The focus on phonetic detail is also evident in Prom-on et al. (2016), who compared intonation models by synthesizing pitch contours according to each model's principles, calculated the root mean square error (RMSE) between the natural and

[3] IPO modelling does address certain types of variability: pitch tracks are replaced by synthesized versions using straight lines, a process known as stylization. Stylization stops when listeners no longer hear the original and stylized version as the same. Importantly, listeners do not have to specify whether any differences they report are meaningful.

synthesized versions, and rated models with lower RSMEs more successful even if listeners found them as natural as models with higher RSMEs. Such approaches can successfully model F0 but, as argued in Arvaniti & Ladd (2009, 2015) and Pierrehumbert (2022), by focusing on F0 granularity, they may fail to capture significant generalizations about the intonation system they describe.

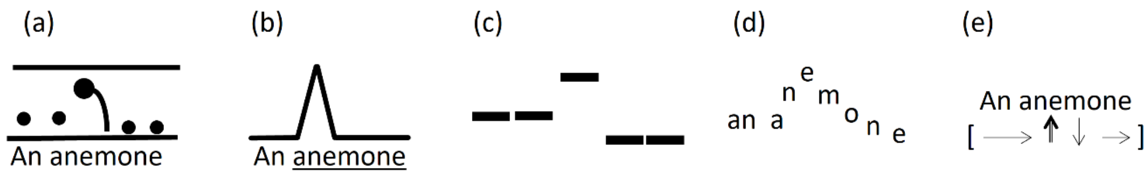


FIGURE 3. Representation of the tune shown in Figures 1b and 2b in the notation of (a) the British School, (b) IPO, (c) the American structuralists; (d) Bolinger, and (e) INTSINT; for details, see text.

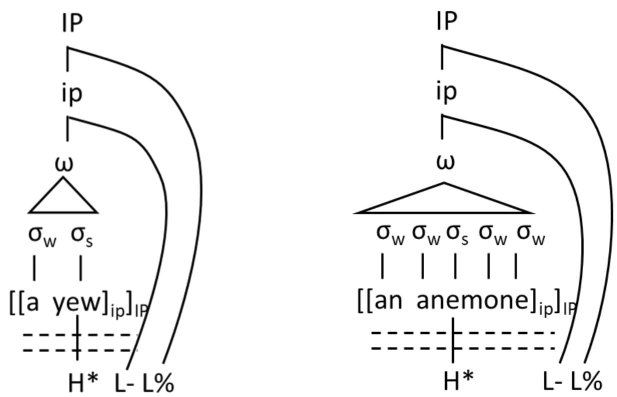
In conclusion, the models briefly reviewed above, whether they focus on iconic representations or F0 approximations, aim primarily to represent F0 with only marginal forays into intonational phonology (for a discussion see Ladd 2008, chap. 2). Nevertheless, most approaches do need to make use of phonological concepts: Fujisaki (1983) originally used accent commands to account for F0 changes related to lexical pitch accent in Japanese, IPO distinguishes between prominence-lending and non-prominence lending pitch movements (’t Hart et al. 1990), while analyses within the British school rely on stress and largely agree on a break-down of tunes into four components, pre-head, head, nucleus and tail (Nolan 2022). An explicitly phonological approach is AM, to which we turn next as it is a dominant model of intonation and does address some of the questions raised in section 1.

1.2 THE TREATMENT OF VARIABILITY IN THE AUTOSEGMENTAL-METRICAL THEORY OF INTONATIONAL PHONOLOGY

AM is a PHONOLOGICAL model in which tunes are broken down into autosegments that represent two types of tonal events, PITCH ACCENTS and EDGE TONES; in most versions of AM, the latter are further distinguished into PHRASE ACCENTS and BOUNDARY TONES. These events associate with structural positions in the metrical tree: pitch accents associate with metrical heads (informally, stressed syllables) and edge tones with the boundaries of two phrase levels, the intermediate phrase or ip, and the intonational phrase or IP (Beckman & Pierrehumbert 1986, Pierrehumbert & Beckman 1988). Pitch accents and edge tones are represented as H (high) and L (low) tones or combinations thereof. Since only metrical heads and boundaries are associated with tones, Ls and Hs are not intended to represent every F0 modulation (Jun 2022).

Based on the above, the pitch contours in Figure 1 have the same autosegmental representation: they both consist of a H* pitch accent associated with the metrical head of each utterance, the stressed syllable of *yew* and *anemone*, followed by a L- phrase accent and a L% boundary tone which associate with the right ip and IP boundary respectively. This is depicted in (1) which shows the autosegmental representation of the tune and a (simplified) representation of the tune's association with metrical structure, following Pierrehumbert & Beckman (1988). By combining the same H* L-L% autosegmental representation with each utterance's metrical structure, AM captures the generalization that the two contours are instances of the same abstract tune. Additionally, their differing METRICAL structures explain and predict how, where, and why the actual F0 contours, the realizations of that tune, will differ between *a yew* and *an anemone*.

(1)



Specifically, AM provides an account of the phonetics of intonation, the gist of which is that intonational categories ALIGN (are synchronized) with the tone bearing units (TBUs) that reflect each category's phonological association (Pierrehumbert & Beckman 1988, chap. 7). In our example, H* is a pitch accent, and as such it is realized as rising F0 on the stressed syllable of the accented word, /nɛ/ in *anemone* and /ju/ in *yew*. The final L% boundary tone is realized as low F0 on the last TBU of the utterance, while the L- phrase accent is realized adjacent to it and spreads left to cover the stretch between the pitch accent and L% (cf. Barnes et al. 2010). The difference in utterance length and stressed syllable location coupled with the requirement that L% reach the speaker's habitual low (Liberman & Pierrehumbert 1984), create the observed differences. In *an anemone* there are two unstressed syllables (-*mone*) on which to reach that low, so the accented vowel is rising and F0 starts to fall after it, reaching a low stretch on the last vowel. In *a yew*, on the other hand, we observe TONAL CROWDING, the presence of more tones than TBUs with which the tones can co-occur, a phenomenon long documented to lead to contour modifications (Bruce 1977, Arvaniti, Ladd & Mennen 2006a, Arvaniti & Ladd 2009, Katsika 2016, Rathcke 2017). In *a yew*, the accented vowel is the last in the utterance and thus it must accommodate both the F0 rise for the H* and the subsequent fall for L-L%; consequently, the accented vowel is rising-falling and the fall does not end in a low F0 stretch but a low F0 point.

As the above example illustrates, AM, by separating phonological representations from phonetic modelling, can capture significant generalizations and predict the form tunes will take when used with different segmental material and diverse metrical structures (Arvaniti & Ladd

2009, Arvaniti 2022a). Thus, it successfully addresses questions A and B above. Although this much is clear, AM still has difficulty handling variability, and distinguishing gradience from noise and linguistically driven variation.

Several reasons have contributed to this state of affairs. First, although AM is a phonological model, its representations are often treated as phonetic transcriptions. Thus, Ls and Hs are seen not as symbolic representations of tonal events but as a means of reconstructing and discretizing F0, much like the F0 idealizations discussed in section 1.1. One could, for example, replace the up and down arrows of INTSINT with H and L respectively and get a rudimentary equivalence between INTSINT and AM, though the similarities are not intentional and the symbols are not meant to have the same status in the two approaches.

The practice of using AM as a phonetic transcription of F0 is perhaps reinforced by the seeming phonetic transparency of the labels L and H which are expected to represent actual low and high F0 points in a contour (for a discussion, see Dilley & Breen 2022 and Jun 2022). Such points, referred to as local F0 minima and maxima or TURNING POINTS, are often the focus of attention during annotation and subsequent analysis (Arvaniti 2022a for a review; for a recent method focusing on such points, see Ahn et al. 2021). Note that this expectation of transparency does not apply to symbols for vowels and consonants and their relation to phonetic substance: the relation between, say, [a] and its phonetic value is recognized to be one of convention, and [a] can be used to represent several vowel qualities (International Phonetic Association 1999). Inevitably, the expectation of transparency in intonation representations poses the same problems faced by all attempts to create phonetically transparent yet discretized symbolic representations, since it is unclear how fine-grained such representations should be (Browman & Goldstein 1992, Ladd 2014, ch.2; for a discussion of this issue in intonation, see Arvaniti 2016).

Most importantly, perhaps, AM has difficulty dealing with many aspects of variability, because phonetic invariance is used in many AM studies as the primary diagnostic criterion of tonehood: invariance is said to apply to TONAL TARGETS, F0 points defined by their scaling (F0 height) and their alignment with the segmental string (Ladd 2008, Arvaniti 2022a). The use of tonal target invariance as a criterion is directly related to the notion of SEGMENTAL ANCHORING, first broached in Arvaniti, Ladd & Mennen (1998) and further developed in Ladd, Mennen & Schepman (2000) and subsequent publications by Ladd and his colleagues. Segmental anchoring posits that tonal targets reflecting underlying tones align with the segmental string in a stable manner. This postulate is based on the finding of Arvaniti et al. (1998) that prenuclear rising accents in Greek align just so: the rise onset aligns with the onset of the accented syllable, while the peak is realized just after the onset of the first postaccentual vowel. Arvaniti et al. (1998) argued that this stable alignment constitutes evidence of two tonal targets, L and H, leading to the Greek accent in question being represented as L*+H.

In the past 25 years, segmental anchoring has been extensively used as a diagnostic criterion for phonological status: F0 points that show stable phonetic alignment with segmental landmarks are treated as tonal targets and thus as reflexes of underlying tones; by extension, more variable turning points are considered epiphenomenal (see Jun 2005, 2014 for illustrations). Stylized synchronization of F0 with syllables is found in several publications and the depicted differences in tonal alignment are said to reflect distinct phonological categories (e.g. Arvaniti & Garding 2007 on American English, Prieto 2014 on Catalan, Frota 2014 on European Portuguese).

An unintended corollary of segmental anchoring has been the implicit expectation that tonal targets are invariant (rather than simply stable), remain distinct across tonal events, and show minimal, if any, cross-category overlap. This is necessary for the invariance criterion to work. For instance, Prieto (2014) posits L+H* and L+<H* accents for Catalan based on the location of the peak, a difference depicted in Figure 4. Hualde & Prieto (2016) propose that these two peak alignments be adopted cross-linguistically. Both proposals assume that the differences can be contrastive in a system. For such precision to work in practice, however, tonal targets must be invariant or very close to it, because even a small deviation from these idealizations would lead to mis-categorization.

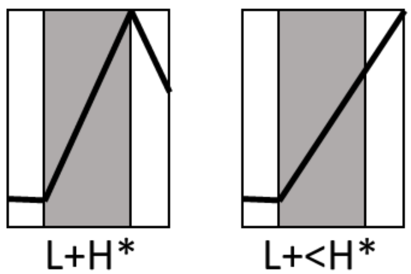


FIGURE 4. Representations of the alignment contrast between L+H* and L+<H* accents, after Prieto (2014) and Hualde & Prieto (2016); thick black lines represent idealized F0; gray rectangles represent the stressed syllable.

The assumption of tonal target invariance is seemingly supported by several studies. However, as Arvaniti (2016) observes, intonation studies often rely on scripted data elicited from dialectally homogeneous groups of educated speakers who are at ease with reading from a script in a consistent manner; consequently, the data of many studies is relatively uniform. The trend of relying on scripted data for intonation research does show signs of waning; for example, Prieto (2014), Holliday (2019) and Baltazani et al. (2022) rely at least in part on spontaneous speech. Nevertheless, the reliance on scripted data applies to the bulk of the existing literature, including both seminal papers (e.g. Liberman and Pierrehumbert 1984, Pierrehumbert & Beckman 1988, Silverman & Pierrehumbert 1990, Arvaniti et al. 1998) and recent research (e.g. Jun & Cha 2015, Grice, Vella & Bruggeman 2019, Prieto & Roseano 2021).

In short, the above criteria and practices lead to an expectation that phonetic features such as the early vs. late peak alignment distinguishing L+H* and L+<H* in Figure 4 have distributions similar to those in Figure 5a: narrow, indicating minimal variability, and entirely distinct from each other, reflecting a complete lack of category overlap. Invariance and lack of category overlap may be desirable for annotators wishing to classify tonal events based on phonetic

criteria, but they are at odds with the extent of variability documented in natural speech, which is closer to the overlapping distributions depicted in Figure 5b.

Variability and category overlap are not new findings in phonetics. They have been known for decades and have been the topic of extensive debates (Pierrehumbert 2002, 2016, Scobbie 2007, Ladd 2014, chap. 2, inter alia). Category overlap is perhaps best known to apply to vowels. Though it is often assumed to be the consequence of differences in vocal tract length, overlap persists even after normalization, as illustrated in Figure 5c with data from male English speakers from New York (Strange et al. 2007). Strange et al. (2007) used linear discriminant analysis to classify vowels of New York English, Parisian French, and Northern German produced in careful speech and report that some vowels are correctly classified just 50% of the time. Category overlap does not apply only to vowels. A case in point is English VOT which reflects a robust phonological contrast. Nakai and Scobbie (2016) applied rate-independent optimal VOT category boundaries to homorganic voiced and voiceless stops and found significant overlap, ranging from 3.3% to 18.3% (see Figure 5d). Overlap is also present in the timing of tone and lexical pitch accents (e.g. Karlin 2018, 2022). These examples show that invariance and lack of category overlap are illusions. Yet, realistic distributions, like those depicted in Fig. 5b, are considered problematic for intonation categories because they evince overlap.

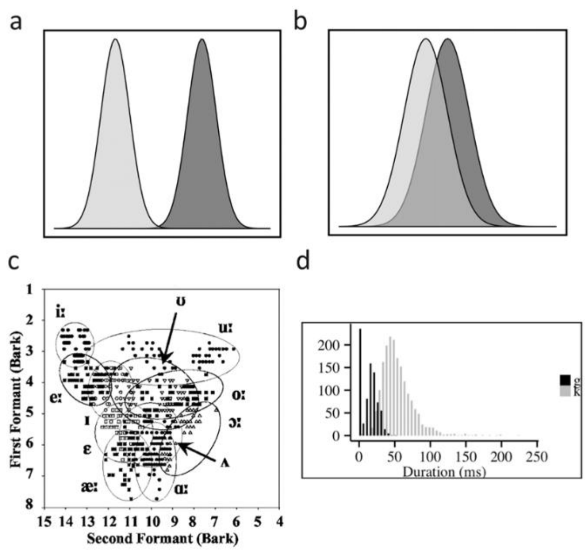


FIGURE 5. (a) idealized distributions of early vs. late peak alignment; (b) realistic distributions of the putative distinction depicted in (a); (c) F1 × F2 plot of American English vowels as produced by New York males (Strange et al. 2007); (d) VOT distributions for English /g/ and /k/ in spontaneous speech (Nakai & Scobbie, 2016).

The implicit connection between segmental anchoring, invariance and lack of category overlap is not only observationally inadequate, it is also problematic for two additional reasons. First, invariance is difficult to statistically confirm. Second, the use of invariance as a criterion is circular: categories are said to be distinct if their tonal targets align differently; in turn, observed differences in target alignment are said to reflect distinct categories. Circularity can only be avoided if posited categories are based on INDEPENDENT EVIDENCE, such as the use of pragmatic criteria. This is a point to which we return below.

To sum up, the approaches to intonation so far have either ignored variability or made it into a core issue: it is either a focus of modelling or a problem to be solved. This is largely due to the fact that, despite some forays into phonology, most approaches focus on modelling F0 rather than intonation per se, or do not make a principled distinction between the two. AM gets closer to explaining linguistically-controlled variation, but cannot address – in its current form and with prevailing practices – other aspects of variability, especially gradience.

1.3 CUE ENHANCEMENT AND CUE TRADING IN INTONATION

The above discussion focused on F0, since most approaches implicitly assume F0 to be intonation's sole phonetic exponent. Recent studies, however, report interactions between F0 and parameters that pertain to segments (such interactions apply to lexical uses of F0 as well, as discussed in section 4). In German, fricative spectral energy is affected by the choice of a rising vs. a falling tune (Niebuhr 2012). In Greek, vowels are longer when co-produced with accentual than phrasal rise-falls (Arvaniti, Ladd & Mennen 2006b), while voice quality varies systematically with accent type (Hu & Arvaniti 2023). In Polish, vowels co-produced with the urgent calling tune are louder and longer than those produced with the routine calling tune (Arvaniti, Żygis & Jaskuła 2017). Articulatory studies also provide evidence for co-modulations of F0 and articulatory kinematics. In German accented syllables, constriction formation duration and articulatory displacement increase along with F0 across focus types, from broad to narrow to contrastive focus (Roessig & Mücke 2019). Similarly, in Greek, both the onset of boundary-induced lengthening of articulatory constrictions and the onset of F0 movements corresponding to boundary tones are timed with respect to the position of the last stressed syllable in the utterance (Katsika et al. 2014, Katsika 2016).

In the above-mentioned literature, the focus has been on the effect of intonation on segments (Roettger & Grice 2019 for a review). However, it is also possible for these connections to lead to *enhancement* of intonation categories instead: If a non-tonal change boosts some tonal feature, the former can become stably associated with an intonation category that relies on that feature, thereby becoming a secondary element in that category's realization. Perceptual integration, such as that reported by Steffman and Jun (2019) for F0 and duration, points in this direction. Enhancement is known to routinely apply to segments and leads to more stable category differentiation (cf. Stevens & Keyser 1989); for example, rounded back vowels are more frequent than unrounded ones because lip rounding lowers F2, thereby enhancing a feature that already distinguishes back from front vowels.

In addition to enhancement, it is possible that non-tonal cues are used to counter the undershooting of tonal features. In other words, the relationship may become one of *cue trading*, whereby when one cue decreases, the other increases in turn. Like enhancement, cue trading is not a new concept when it comes to segments (see Schertz & Clare 2020 for a review). For instance, although English voiced and voiceless stops are distinguished primarily on the basis of VOT, closure duration and F0 at the onset of voicing also contribute to their differentiation and can be recruited to do so if the appropriate circumstances arise (Repp 1982). The notion of cue trading has yet to be directly assessed in intonation. Here, we take a first step toward doing so by examining the relationship between F0 and duration to better understand whether it lends itself to a cue enhancement or a cue-trading interpretation.

1.4 Outline

To address the research questions discussed above, we used Functional Principal Components Analysis (henceforth FPCA; Ramsay & Silverman 2005) in a corpus of Greek data designed to elicit three accents represented as H*, L+H* and H*+L in AM (Arvaniti & Baltazani 2005). In the remainder of the paper, we first present the three accents, explain the workings of FPCA, and discuss its output in relation to the accents, focusing on between-category overlap. This analysis is followed by FPCA of the L+H* F0 curves used to examine context-dependent within-category variation. Finally, we use FPCA that combines F0 and duration to examine their joint effects on the realization of L+H* and thus explore the issue of cue-trading. Our findings indicate that dimension-reduction methods like FPCA can be useful in handling pervasive variability in intonation but are also sufficiently sensitive to explain linguistically determined variation. We use these findings to present a conceptualization of intonation that addresses the main questions posited in the introduction. We adopt an AM understanding for our analysis but stress that our findings are relevant for anyone working on intonation (or lexical tonal phenomena, for that matter), independently of framework adopted, as the method is data-driven and the results shed light on the nature of intonational categories and the methods by which they can best be investigated.

2 Methods

2.1 Speakers and procedure

The data was elicited from 13 native speakers of Standard Greek (10 F, 3 M), aged from early 20s to early 40s (mean = 34, S.D. = 8). The speakers lived in Athens and its suburbs at the time of the recordings. None reported a history of speech or hearing disorders.

The participants were recorded in quiet locations using a DAT recorder at a sampling rate of 44.1 kHz. The short dialogues used to elicit the data were typed on cards in Greek orthography. The participants read the dialogues four times, with the cards being shuffled between repetitions, as part of a larger dataset with different utterance types acting as distractors.

2.2 Greek H*, L+H* and H*+L

According to GRToBI (Arvaniti & Baltazani 2005), in Greek, H*, L+H* and H*+L – illustrated in Figure 6 – are pitch accents used in phrase-final (i.e. nuclear) position in declaratives. In this position, they are followed by L-L% edge tones, viz. H* L-L%, L+H* L-L%, H*+L L-L. Each accent is used in different pragmatic situations.

H* is the default final accent in broad focus declaratives. It indicates that the accented item is new in discourse and should become part of the common ground (Arvaniti & Baltazani 2005; cf. Pierrehumbert & Hirschberg 1990). For example, in Figure 6a, H* is used on [ˈmavro] 'black' which represents new information in response to a question such as 'what color is your new sweater?'. Phonetically, H* creates a plateau with any preceding accentual peak, if one is present; if H* is the only accent, as in Figure 6a, it is realized as a (minimal rise-)fall (Arvaniti & Baltazani 2005).

L+H* indicates narrow focus which can be corrective or contrastive (Arvaniti & Baltazani 2005, Arvaniti et al. 2006b, Georgakopoulos & Skopeteas 2010). It indicates that the accented item should be chosen from amongst a closed set of (explicitly mentioned or easily inferable) alternatives and added to the common ground. For example, in Figure 6b, L+H* is used on [ˈmavro] 'black' in response to a question such as 'is your sweater black or navy?'. L+H* is realized as a rise to a peak; the rising shape is retained in the presence of a preceding H tone (Arvaniti & Baltazani 2005, Arvaniti et al. 2006b).

H*+L encodes new information, like H*, but indicates that the speaker believes it should have already been part of the common ground (Arvaniti & Baltazani 2005). Thus, H*+L can be used to implicate that one is stating the obvious or is being asked to provide information they consider redundant; depending on context, it can reflect surprise or impatience. For example, in Figure 6c, H*+L is used on [ˈmavro] 'black' in response to a question such as 'what color are you wearing at the funeral?', a redundant question, since the only socially appropriate color to wear at a Greek funeral is black. H*+L is realized as a high fall; its high scaling is evident in Figure 6c in which its' peak is approximately 100 Hz higher than that of H* and L+H* (see also Figure 4.3 in Arvaniti & Baltazani 2005).

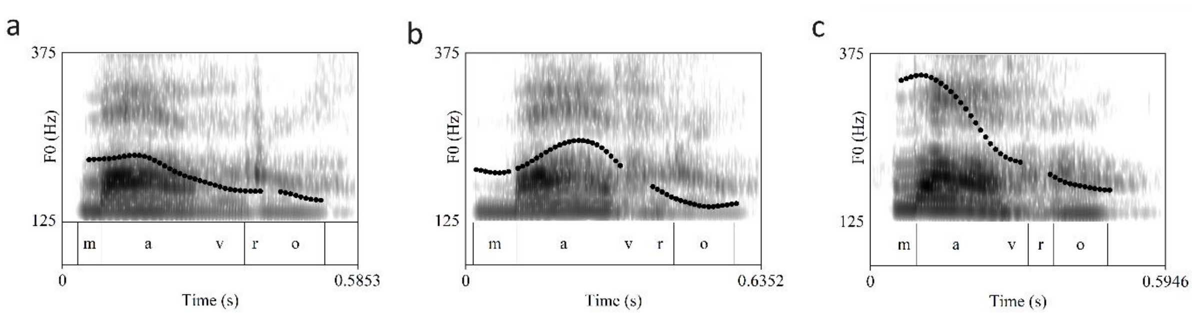


FIGURE 6. H* (a), L+H* (b), and H*+L (c), as realized on a single-word utterance [ˈmavro] 'black' produced by the same female speaker during the same recording session.

These three accents are an ideal set of categories to investigate intonation variability. First, because of their distinct pragmatics they must be elicited using different prompts; this avoids circularity in analysis, since accent classification can be based on the pragmatics of the prompts, rather than the phonetic realization of the accents. Second, the accents are phonetically comparable but also highly susceptible to variation due to tonal crowding: Greek restricts the location of stress to the last three syllables of a word, and since these accents appear on the utterance-final word, the combination of the location of the accent and the stress restrictions means that the accent and following edge tones (H* L-L%, L+H* L-L%, H*+L L-L%) must all be realized on at most three syllables, and may have to be realized on one syllable only. As noted in section 1.2, tonal crowding alters tonal realization: languages make adjustments to accommodate the tune, such as tone truncation, realignment (e.g. bringing tones forward), compression, and undershoot (Roettger & Grice 2019). These effects can add to within-category variability which in turn can lead to cross-category overlap.

2.3 Materials

The analysis is based on the test words shown in (2), stressed on the antepenult (2a), penult (2b) and ultima (2c).[4] This change of stress location was selected to manipulate the degree of tonal crowing, that is, the pressure from the edge tones as the stressed syllable moves closer to the end of the utterance. The test words were of different length to ensure that each had two unstressed syllables preceding the stressed one.

(2) a. [laðoˈ**le**mono]#
‘*oil-and-lemon sauce’*
b. [lemoˈ**na**ða]#
‘*lemonade’*
c. [ɣalaˈ**na**]#
‘*light blue’*

The test words were final in the answers of short Q&A dialogues. Representative dialogues are shown in (3) together with the expected tune and its association with the segmental string (for the full set, see https://osf.io/sjxuq/?view_only=985aa03025524b71b8a0dbc801d7a164). Answers consisted of either one or two content words. Answers with one content word had only one accent, that of the test word, as shown in (3a-c). Answers with two content words had two accents, a prenuclear L*+H on the first word and the target accent on the test word; thus, the target accents were always utterance-final nuclear accents, as shown in (3d-f). The prenuclear accent was always separated from the nuclear accent by three unaccented syllables to minimize tonal pressure from the former on the latter. Henceforth, we refer to the difference between utterances with one or two accents as Accentual Context (present when a prenuclear accent precedes, and absent when it does not).

In total, 936 tokens were elicited [13 speakers × 3 accents × 3 test words × 2 utterance lengths × 4 repetitions], of which 844 were used for analysis, 272 tokens of H*, 274 of H*+L, and 298 of L+H*. The participants had no difficulty producing contours with the intended accent and focus structure. Nevertheless, 92 tokens had to be discarded because of background noise, disfluencies, and extensive stretches of creak which led to unreliable F0 tracking; most of these

[4] The word [laðoˈlemono] is a compound. Compounds in Greek are single prosodic words with only one stressed syllable and thus they carry only one accent (Nespor & Ralli 1996).

tokens were elicited from speaker F6 so that only her L+H* contours could be retained. This does not materially affect the analysis. As mentioned in section 2.2., in order to avoid circularity and bias, whenever classification into linguistic categories was required during analysis, tokens were classified into one of the three accentual categories based on the dialogue in which they were elicited, not their phonetic realization; in other words, classification was not based on accent shape but on pragmatic function as determined by the context. To keep the accent descriptions short, in the following sections we use the AM representations (H*, L+H*, H*+L) as a shorthand for the three accent categories. In section 3.2 we discuss these representations in light of our results.

(3) Accentual Context: Absent

H*

a. Q. *What's this?*

A. [[laðoˈlemono]ip]IP

    | | |
    H*   L- L%

Oil-and-lemon [sauce].

H*+L

b. Q. *What should I make with so many lemons?*

A. [[lemoˈnaða]ip]IP

    | | |
    H*+L- L%

Lemonade.

L+H*

c. Q. *Did you say their son's eyes are brown?*

A. [[ɣalaˈna]ip]IP

    | | |
    L+H* L-L%

Blue! [Aren't you listening?]

Accentual Context: Present

H*

d. Q. *What will you have?*

A. [[sinaˈɣriða laðoˈlemono]ip]IP

    | | | |
    L*+H   H*   L- L%

Seabream in oil-and-lemon [sauce].

H*+L

e. Q. *What should I do with all these lemons?*

A. [[ˈkane lemoˈnaða]ip]IP

    | | | |
    L*+H   H*+L - L%

Make lemonade.

L+H*

f. Q. *Are their daughter's eyes brown?*

A. [[ˈine ɣalaˈna]ip]IP

    | | | |
    L*+H L+H* L-L%

They're blue! [Don't you remember her?]

2.4 MEASUREMENTS AND THE ISSUE OF VARIABILITY

As an illustration of the basic shape of each accent in the corpus, the smoothed average contour of each accent in normalized time and F0 range is shown in Figure 7. These normalized averages mask substantial variability, as illustrated in Figure 8, which presents smoothed and normalized averages by speaker and separately for one- and two-word utterances. Although, by and large, all speakers differentiated the accents in both contexts, it is clear they did not do so in identical fashion or to the same extent. Assuming that the aim of any analysis, independently of theoretical framework, is to capture the differences between accents, this variability can create difficulties during analysis. These were the difficulties we aimed to address by using FPCA.

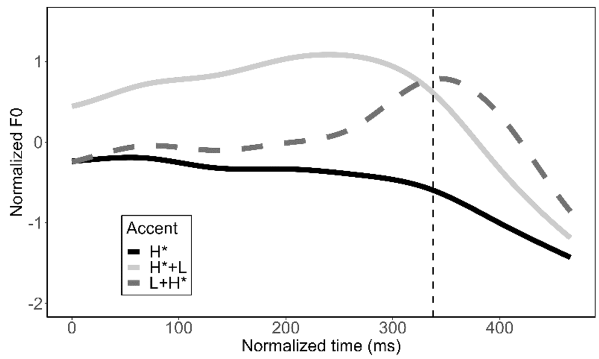


FIGURE 7. Smoothed and normalized average F0 curves per accent category, pooled across stress, accentual contexts, and speakers; the figure shows the course of F0 on the first two syllables of the test words plus the accented syllable, viz. [laðoˈle] in [laðoˈlemono], [lemoˈna] in [lemoˈnaða] and [ɣalaˈna] in [ɣalaˈna]; the broken vertical line indicates the onset of the accented vowel.

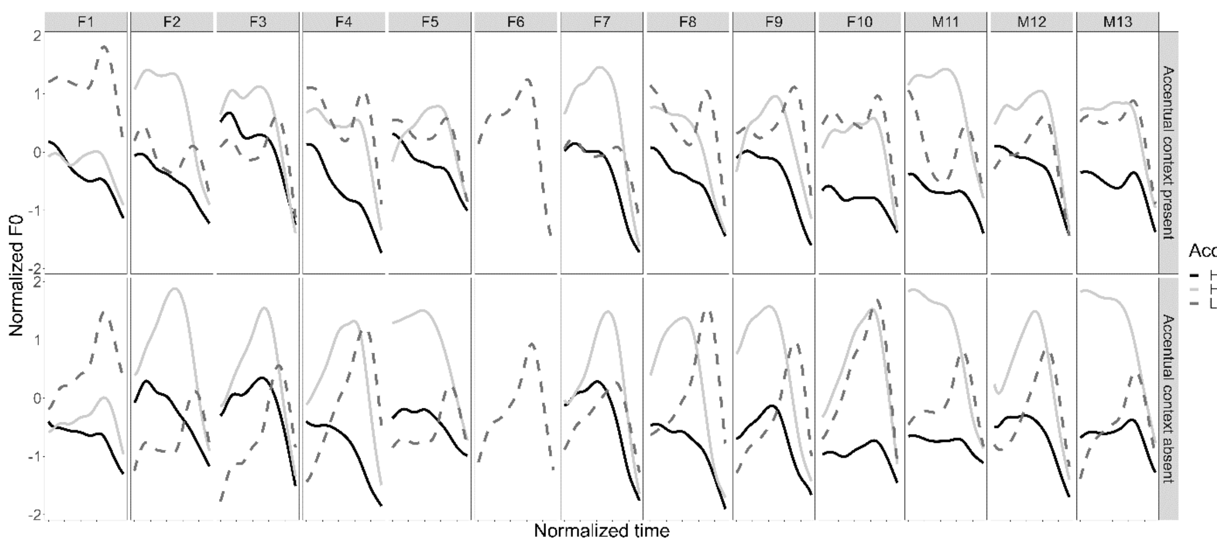


FIGURE 8. Smoothed and normalized average F0 curves per accent category, separately for each speaker and accentual context (note that F6 provided only L+H* curves); top: accents from two-word utterances in which the target accents were preceded by a prenuclear L*+H; bottom: accents from one-word utterances, that is without a preceding accent. The data include the first three syllables of the test words, viz. [laðoˈle], [lemoˈna], [ɣalaˈna].

2.5 FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS

FPCA is a data-driven, dimension-reduction form of Functional Data Analysis (FDA), in which sampled curves are interpolated and represented as continuous functions of time, before their dominant and independent modes of variation, called PRINCIPAL COMPONENTS or PCs, are extracted (Ramsay & Silverman 2005; see Gubian, Torreira & Boves 2015 and Asano & Gubian 2018 for details of FPCA use in the study of F0 and intonation; see Grabe, Kochanski & Coleman 2007, Chen and Boves 2018, and Baltazani et al. 2022, for uses of FDA in the study of intonation).

In FPCA, the shape of each F0 curve in the sample is modelled by Eq. (1): $f(t)$ represents the shape of each F0 curve, $\mu(t)$ is the mean curve of all input curves, and $PC1(t)$, $PC2(t)$, etc. are the principal component curves, each associated with a coefficient or SCORE ($s1$, $s2$, etc.) that represents the PC's contribution to the shape of the corresponding F0 curve.

Eq. (1) $f(t) \approx \mu(t) + s1 \times PC1(t) + s2 \times PC2(t) + \ldots$

The PCs are not actual curves in the data, but curve components that collectively determine the shape of each input curve. The PC scores can be used to reconstruct curves, as shown in 3.1, and can also be subjected to statistical analysis to determine whether they differ depending on linguistic categories under consideration; for example, in our analysis of the entire dataset, the linguistic categories were the three accents. For our research purposes, FPCA offers three advantages relative to a frequently employed alternative, Generalized Additive Models (GAMs). First, FPCA decomposes variance into independent components allowing us to test claims related to the main features of tonal events, such as the AM concepts of scaling and alignment (see section 3.1). Further, FPCA does not require a priori classification of curves into linguistic categories, and thus, in the first stage of the analysis at least, the output is independent of linguistic classification. Finally, as shown in section 3.5, FPCA allows us to jointly consider the contribution of F0 and duration to the realization of intonation categories, something that is not as yet possible with GAMs.

We followed the methodology of Gubian et al. (2015) with some modifications (Gubian et al.'s scripts can be found at https://github.com/uasolo/FPCA-phonetics-workshop; scripts and outputs used in the present study can be found at: https://osf.io/sjxuq/?view_only=985aa03025524b71b8a0dbc801d7a164). We applied FPCA to a three-syllable window consisting of the accented syllable of each target word and the two unstressed syllables preceding it, viz. [laðoˈle] in [laðoˈlemono], [lemoˈna] in [lemoˈnaða] and [ɣalaˈna] in [ɣalaˈna]. This window was selected because it captured F0 variation we had observed on the unstressed syllables and was comparable across target words: it always consisted of two unstressed and one accented syllable and captured degrees of tonal crowding due to the accented syllable getting increasingly closer to the end of the utterance. The onset and offset of the three-syllable window and the onset of the accented vowel were annotated in Praat (Boersma & Weenink 2018), following standard criteria of segmentation (Machač & Skarnitzl 2009).

The F0 of the analysis window was extracted using STRAIGHT (Kawahara et al. 2005) in VoiceSauce, which estimates an F0 value every 1 ms throughout the utterance, with a moving window length of 25 ms (Shue 2010, Shue et al. 2011). The VoiceSauce default F0 range setting of 40-500 Hz was used.[5] F0 was measured in Hz and normalized by speaker, by using the scale function in base R, which converts the values into z-scores by centering and scaling (i.e. first, the mean is subtracted from each value and then the values are divided by the standard deviation). In Gubian et al. (2015) and Asano & Gubian (2018), normalization was done by curve, that is, by subtracting each point of each curve from the curve's average. We did not follow this procedure because it eliminated pitch level differences between accents which we knew by observation to be present (cf. Figures 6, 7 and 8).

F0 normalization was followed by a step known in FDA as LANDMARK REGISTRATION (Ramsay 2005), which warps time by fitting each curve to the same interval of normalized time and aligning all curves around the landmark, a specific point in the analysis window. The duration of the landmark-registered curves is the mean duration of all curves; thus, duration ranges from 0 to the mean, which is presented in ms in relevant figures. The process is illustrated in Figure 9. Here, the selected landmark was the acoustic onset of the accented vowel. This separated the analysis window into two intervals: (i) the two unaccented syllables plus the accented syllable onset, and (ii) the accented vowel; e.g. [ɣalaˈna] was split into [ɣalan] and [a]. This choice reflects the AM understanding that accents phonetically align with the vowel of the syllable they phonologically associate with and made it possible to observe changes in the accented vowel's duration, since this vowel was the last interval in the analysis window.

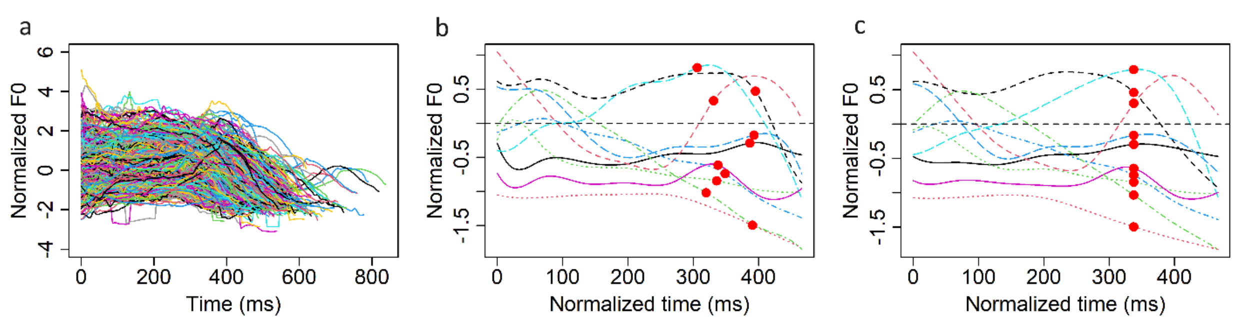


FIGURE 9. In (a), the raw F0 contours of the entire corpus; in (b), a subset of these curves after smoothing (when durations are normalized), but before landmark registration; the red dots indicate the onset of the accented vowel in each curve; in (c), the same curves after landmark registration, with the accented vowel onsets lined up.

The speaker-normalized and landmark-registered F0 curves were smoothed using k = 8 and λ =10000: k refers to the number of *knots*, which corresponds to the time resolution of the model, while λ values indicate the extent to which smoothing is penalized. The values of k and λ were selected from a number of possible combinations after comparing a randomly selected set of smoothed and original curves. The selected values were chosen to reflect the input, while avoiding both over-smoothing, which could eliminate important curve changes, and over-fitting, which could lead to modelling microprosodic effects. After smoothing, the curves were

[5] This range was deemed appropriate, as utterances with extensive creak that could have led to pitch tracking errors were excluded before analysis. Eight tracks (out of 844) that upon visual inspection presented some errors were retained, since FPCA is robust and we did not wish to further decrease sample size.

submitted to FPCA, executed using the fda package (version 5.1.9; Ramsay, Graves, & Hooker 2020).

3 RESULTS

3.1 FPCA OF F0 CURVES AND STATISTICAL MODELLING BY ACCENTUAL CATEGORY

FPCA of the 844 F0 curves in the corpus showed that the first three PCs explained 96.4% of curve variance (see Table 1). The shape of these PCs is illustrated in Figure 10. PC1 reflected primarily differences in scaling: higher PC1 scores (red lines) were likely to result in curves with higher scaling, but also in a somewhat earlier fall; lower PC1 scores (blue lines) resulted in lower scaling and a later fall. Higher PC2 scores resulted in a rise-fall shape with a high and late peak, while low scores resulted in a plateau with a low and early 'peak' (effectively, the end of the plateau). PC3 high scores resulted in a plateau with a late fall and low scores in an early-timed fall after a slight rise.

PC	% of curve variance explained by each PC	R^2m by PC score	R^2c by PC score
PC1	65%	0.34	0.36
PC2	22.6%	0.36	0.68
PC3	8.8%	0.05	0.55
PC4	2.2%	0.08	0.28
PC5	0.9%	0.03	0.15

TABLE 1. Percentage of curve variance explained by each PC resulting from the FPCA of all F0 curves, followed by marginal and conditional R^2 as calculated by the LMER fitted to the scores of each PC.

The shape of the PCs informs us about the dimensions along which the data differ, but the FPCA output cannot, on its own, determine how the variance explained by each PC is related to the accent categories under investigation. Variability has many sources and it is possible that a PC's variance is unrelated to the accents (cf. Gubian et al. 2019). To further test how distinct the accents were in their PC scores, we modelled the scores using linear mixed effects models (LMERs; lme4 package version 1.1.26; Bates et al. 2015) run in R (version 4.0.3; R Core Team 2020), following the procedure below. Finally, the mean PC scores estimated by the LMERs were used for curve reconstruction, as an additional test of the relationship between the modelled accent types and the accent shapes observed in the data.

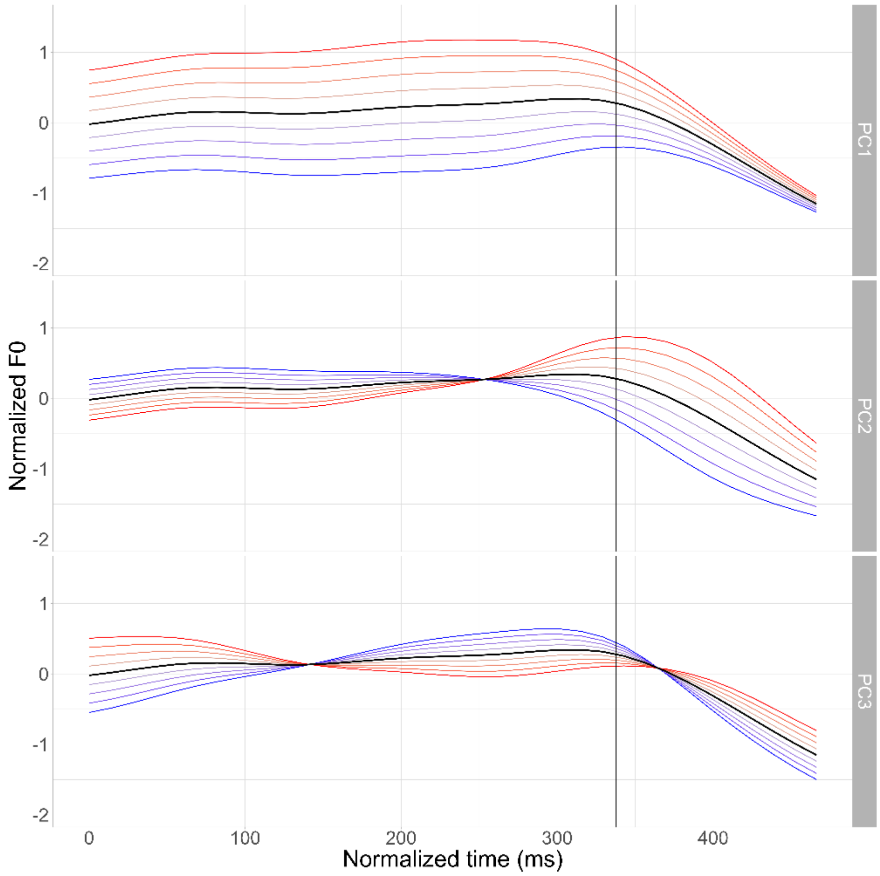


FIGURE 10. FPCA output as applied to the entire dataset of F0 curves. Color-coded curves illustrate the effect of each PC on the mean curve (solid black line) and were obtained by adding to the mean curve a specific PC curve multiplied by a PC score, as per Eq. (1); the equation for the curves in each panel is $f(t) = \mu(t) + \alpha \times \sigma \times (s_n) \times PC_n(t)$, where σ is the standard deviation of the PC coefficient, and α determines the fraction of the standard deviation shown. Alpha ranges from -1 (blue) to +1 (red) in increments of 0.25. The mean curve corresponds to alpha = 0 and is the same in all panels. The vertical line indicates the onset of the accented vowel.

The goal of the statistical modelling of the entire dataset was to examine PC score differences due to accent type. Thus, for each LMER, the scores of one PC at a time were the dependent variable, with Accent (H*, L+H*, and, H*+L) as the fixed effect. H* was the reference level. The full model included random intercepts for both Speaker and Item, with Accent included as a random slope for both Speaker and Item. Item captured differences in the location of stress ([laðoˈlemono] [lemoˈnaða] [ɣalaˈna]) as well as differences in accentual context, viz. the absence or presence of a L*+H prenuclear accent. The purpose of relegating the contextual and stress effects to the random structure (combined under Item) was to focus this analysis on cross-category differences; the contribution of context and stress was assessed in a separate analysis, presented in section 3.3. The p-values were calculated using the lmerTest package (version 3.1.3, Kuznetsova et al. 2017), and variance explained was calculated using the MUMIn package (version 1.46.0, Bartoń 2022). Pairwise comparisons were conducted using the emmeans package (version 1.6.0, Lenth 2021) with the the 'holm' adjustment applied. In all models, REML was set to false.

For each PC, we began with the most complex random structure. In cases of models not converging, we performed reduction, starting from the slope with the lowest variance. Slopes were removed before intercepts in the random structure. Speaker was removed from the random structure of the model for PC1. Optimization conducted with the allFit package (https://joshua-nugent.github.io/allFit/) using the optimizers L-BFGS-B, nlminb, nlm, and bobyqa, did not alter this outcome which, we contend, was due to the fact that F0 was normalized by speaker and thus F0 range differences within each speaker's data were minimized.

Fixed Effects	Estimate	Std. Error	df	t-value	Pr(>|t|
(Intercept)	-11.135	1.193	11.903	-9.338	***
AccentH*+L	23.145	1.104	838.103	20.962	***
AccentL+H*	10.327	1.082	838.149	9.546	***
Random Effects					SD
Item (Intercept)					2.205
Residual					12.898

TABLE 2A. LMER results for PC1 scores from the entire dataset; final formula in R: lmer(PC1 ~ Accent+(1|Item), data = AllAccents_1D, REML=FALSE). Key to significance levels: $p < .05$ = *; $p < .01$ = **; $p < .0001$ = ***.

contrast	estimate	SE	df	t.ratio	p.value
(H*) - (H*+L)	-23.1	1.11	840	-20.936	<.0001
(H*) - (L+H*)	-10.3	1.08	840	-9.534	<.0001
(H*+L) - (L+H*)	12.8	1.08	840	11.853	<.0001

TABLE 2B. Results of the pairwise comparisons within the factor Accent for PC1 scores.

The models for PC1 and PC2 are given in Tables 2a-b, 3a-b; the results are illustrated in Figure 11. Although we run models for the first five PCs, here we present only the results for PC1 and PC2 (for FPCA output including higher PCs and LMERs on their scores, see https://osf.io/sjxuq/?view_only=985aa03025524b71b8a0dbc801d7a164). This decision was based on weighing several pieces of information: the percentage of variance explained by each PC, curve reconstruction, and the estimation of marginal R^2 (R^2m) and conditional R^2 (R^2c) of the LMER for the scores of each PC (Nakagawa & Schielzeth 2013, Johnson 2014). R^2m is an estimate of the fraction of the score variance explained by the fixed effects, while R^2c takes both fixed and random terms into account. Thus, looking at Table 1, we can say that PC2 explained 22.6% of curve variance and that 36% of the variance of its scores was related to accent; this percentage increased to 68% when we also considered random effects, which, as a reminder, included context and stress. In contrast, PC3 explained 8.8% of curve variance which, based on the low R^2m (0.05) and high R^2c (0.55), was unlikely to relate to the accentual contrast. In this analysis, we focused primarily on R^2m, since our aim was to explain accent-related effects and R^2m was high only for PC1 and PC2.

Fixed Effects	Estimate	Std. Error	df	t-value	Pr(>|t|
(Intercept)	-5.952	1.097	12.536	-5.424	***
AccentH*+L	3.792	1.7	10.951	2.231	*
AccentL+H*	13.318	2.271	10.589	5.865	***
Random Effects					SD
Speaker (Intercept)					2.533
Accent: H*+L by Speaker					3.337
Accent: L+H* by Speaker					4.313
Item (intercept)					1.836
Accent: H*+L by Item					3.237
Accent: L+H* by Item					4.568
Residual					5.386

TABLE 3A. LMER results for PC2 scores from the entire dataset; final formula in R: lmer(PC2 ~ Accent + (Accent | Item) + (Accent | Speaker), data = AllAccents_1D, REML=FALSE). Key to significance levels: $p < .05$ = *; $p < .01$ = **; $p < .0001$ = ***.

contrast	estimate	SE	df	t.ratio	p.value
(H*) - (H*+L)	-3.79	1.81	11.9	-2.098	0.0579
(H*) - (L+H*)	-13.32	2.41	11.3	-5.522	0.0005
(H*+L) - (L+H*)	-9.53	1.96	15.6	-4.86	0.0005

TABLE 3B. Results of the pairwise comparisons within the factor of Accent for PC2 scores.

As shown in Table 2a-b, H*+L had significantly higher PC1 scores than both L+H* and H*, while H* had significantly lower scores than the other two accents. Based on the FPCA output in Figure 10, this result suggests that H* had the lowest F0 scaling and H*+L the highest. The differences are illustrated in Figure 11(a) which shows that the PC1 scores for H* were predominantly negative, those for H*+L predominantly positive, and those for L+H* intermediate.

The LMER on PC2 scores indicated that they were significantly higher for L+H* relative to H* and H*+L, between which there was no significant difference (Table 3a-b). As shown in Figure 11(b), the scores of H* and H*+L were largely negative, while those of L+H* were predominantly positive. Based on the FPCA output in Figure 10, this result suggests that L+H* had a scooped rising-falling shape, while H* and H*+L were falls.

The above interpretation of the FPCA and PC scores is supported by the shapes of the reconstructed curves in Figure 12, which are based on PC1 and PC2 only and are strikingly similar to the observed contour averages in Figure 7.

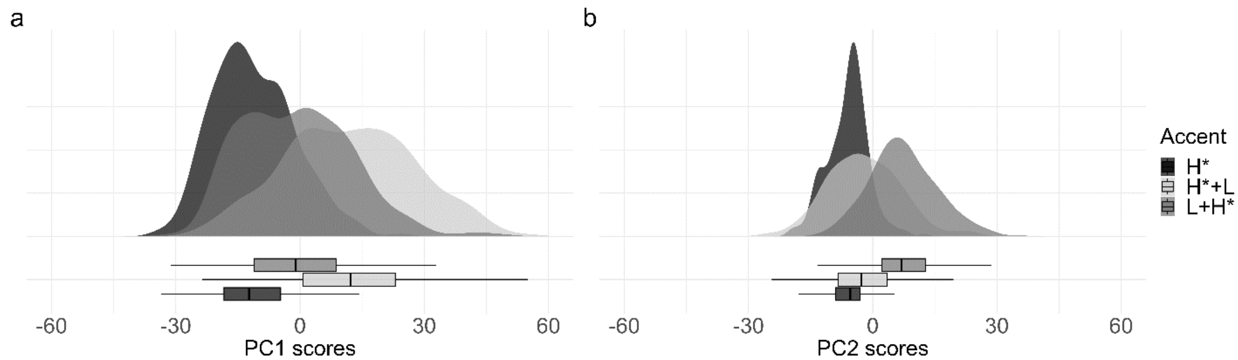


FIGURE 11. Density and box plots of PC1 scores (panel a), and PC2 scores (panel b), by accent (H*, H*+L and L+H*) based on the FPCA of all F0 curves.

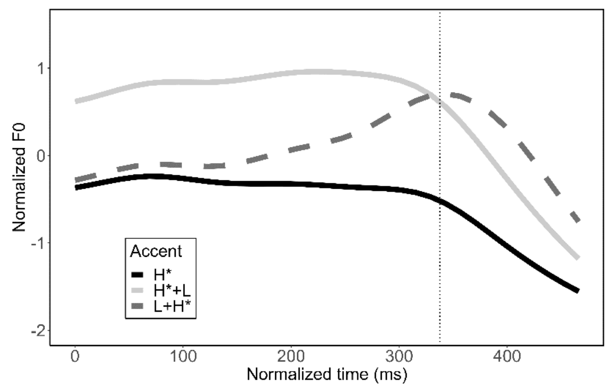


FIGURE 12. Reconstructed F0 curves of the three accents using PC1 and PC2, and obtained by incorporating the average value of each PC score, as predicted by the linear model, into the FPCA equation; Eq. (1): for H*, $f(t) = \mu(t) - 11.1 \times PC1(t) - 5.95 \times PC2(t)$; for L+H*, $f(t) = \mu(t) - 0.80 \times PC1(t) + 7.36 \times PC2(t)$; for H*+L, $f(t) = \mu(t) + 12 \times PC1(t) - 2.15 \times PC2(t)$. The reconstructed curves can be compared to the smoothed, averaged and normalized curves in Figure 7.

Finally, to better understand the distinctions among the three accents, we visualized them in the two-dimensional space defined by PC1 and PC2 scores (Figure 13). We further quantified the accent overlap on that space by computing Pillai scores using the MANOVA() function in R, with the PC1 and PC2 scores as dependent variables and Accent as the independent variable (on Pillai scores, see Kelley and Tucker 2020 and references therein; for details on our analysis, see https://osf.io/sjxuq/?view_only=985aa03025524b71b8a0dbc801d7a164). Pillai scores range from 0 to 1, with values closer to 0 indicating greater overlap. Following Fung and Lee (2019) and Freeman (2021), we interpreted our Pillai scores in conjunction with the statistical significance of the LMER output. As shown in Figure 13, for H* vs. L+H*, the Pillai score was 0.55, and for H*+L vs. L+H*, 0.33; these scores together with the LMER output suggest that these accent pairs were distinct from each other in both dimensions. For H* vs. H*+L, the Pillai score was 0.51, but the LMERs indicate that only their PC1 scores were significantly different, suggesting that H* and H*+L differed largely in scaling but not in shape.

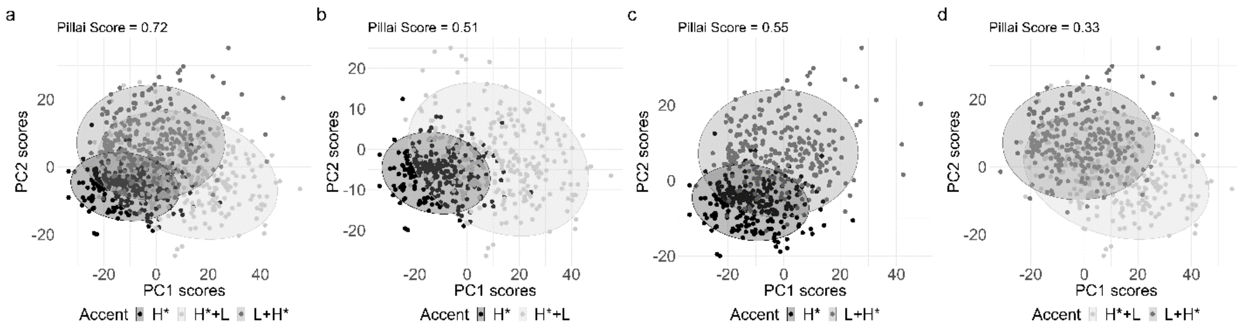


FIGURE 13. Two-dimensional visualizations of H*, H*+L, and L+H* using PC1 and PC2 scores, for all accents together (panel a), and each pair of accents (panels b to d).

3.2 FPCA OF F0 CURVES INTERIM DISCUSSION: ACCENT DIFFERENCES

The results presented above support the analysis posited in GRToBI (Arvaniti & Baltazani 2005) that Greek uses three accents in nuclear position in declaratives. Our analysis showed that their differences are both phonetic and pragmatic: the three accents were used after different prompts to convey distinct information structures and implicatures by means of different phonetic realizations. We thus conclude that they are phonologically contrastive.

A separate question pertains to whether the GRToBI representations of the three accents are also supported by the results. Based on the FPCA output, statistical modelling, and curve reconstructions, we contend that they are. The accent represented as H* was realized as a moderately scaled plateau followed by a gentle fall. Thus, H* seems an appropriate representation for it. The L+H* accent was realized with an F0 dip appearing before the accented vowel and followed by a rise aligned with that vowel. This resulted in a scoop instead of a plateau, a shape that can be plausibly construed as the reflex of a L tone followed by a H tone. Further, since the peak is largely reached on the accented vowel, L+H* is the most appropriate representation (Pierrehumbert 1980, chap. 2). We return briefly to the representation of this accent in section 3.4. Finally, the accent represented as H*+L is realized as a high plateau followed by a steep fall. Thus, H*+L seems appropriate, as it reflects this accent's similarity with H* and its steeper fall, which can be attributed to the presence of three consecutive L tones in this tune, viz. H*+L L-L%. Alternatively, H*+L could be represented as H+L* to reflect the fact that the fall starts before the accented vowel (see Figures 7 and 12), or as ^H* to emphasize the accent's high scaling.[6] H+L* can be dismissed because the accent sounds high, not low (cf. Arvaniti, Ladd, Mennen 2000), while the ^ diacritic for extra high in ^H* introduces phonetic detail into phonological representation, a tendency we argue against (see section 4). We stress, however, that independently of which representation is favored, they all adequately reflect the accentual contrast.

---

[6] We note that such alternatives are not uncommon in phonological representations. The English FACE vowel is represented as a diphthong /eɪ/ in the British tradition but as a monophthong /e/ in the US, while a very similar Dutch vowel is represented as a long /eː/ instead (Gussenhoven 1992). Though the symbols partly reflect phonetic differences, they are also a matter of convention, tradition, and analysis-internal considerations.

Critically, our analysis shows that the three-way contrast can be reduced to just two dimensions: differences in scaling captured by PC1, and differences in shape captured by PC2. PC1 distinguished low-scaled H* from high-scaled H*+L and L+H*, and PC2 distinguished scooped L+H* from plateau-falling H* and H*+L. Note further that some of the differences are distal: the higher scaling of H*+L does not apply to the accented syllable only but to the unstressed syllables preceding it too; it is on those syllables that the rise of L+H* also starts (cf. Figures 7 and 12).

As illustrated in Figures 11 and 13, however, these broad (and statistically significant) differences belie substantial within-category variability and thus cross-category overlap in both dimensions. Nevertheless, Pillai scores in combination with statistical significance (as determined by the LMERs) confirm that the contrast between each accent pair is maintained even when just the PC1 and PC2 dimensions are considered. Additional analyses, not presented here, indicate that the accents are also distinguished by means of duration and voice quality (Hu & Arvaniti 2023; for details, see https://osf.io/qby64/?view_only=1ff260ac07ab4737a261647692dc710a). These differences are a first indication of redundant cues in the realization of intonation categories.

Finally, the LMERs showed that the variance explained by FPCA was not due exclusively to the three accents. A comparison between R^2c and R^2m indicates that at least some variance was due to Item, the random factor that in the LMERs here included differences in stress and accentual context (see Table 1). We turn to a detailed examination of these effects in the next section.

3.3 Effects of stress and accentual context on the L+H* F0 curves

To further examine the sources of linguistically driven variation as they apply within a category, we conducted FPCA on the curves of the L+H* accent (N = 298) followed by statistical modelling of the PC scores and focusing on the effects of stress location and the presence of a preceding accent. We chose L+H* for this analysis because of its shape: its initial F0 dip was likely to show clearer effects related to the presence of a preceding accent than H* and H*+L, while its marked peak made it easier to detect the effect of stress on peak alignment.

For FPCA we followed the procedure described in Section 3.1. Figure 14 shows the first three PCs which together explain 94.2% of curve variance in this dataset (Table 4). Since only L+H* curves were included, all PCs reflected this accent's scooped shape; additionally, they reflected variation in peak height (PC1, PC2), peak location (PC2), and scaling and shape changes to the F0 stretch before the accented vowel (PC1 and PC3 respectively) and the final drop (PC2).

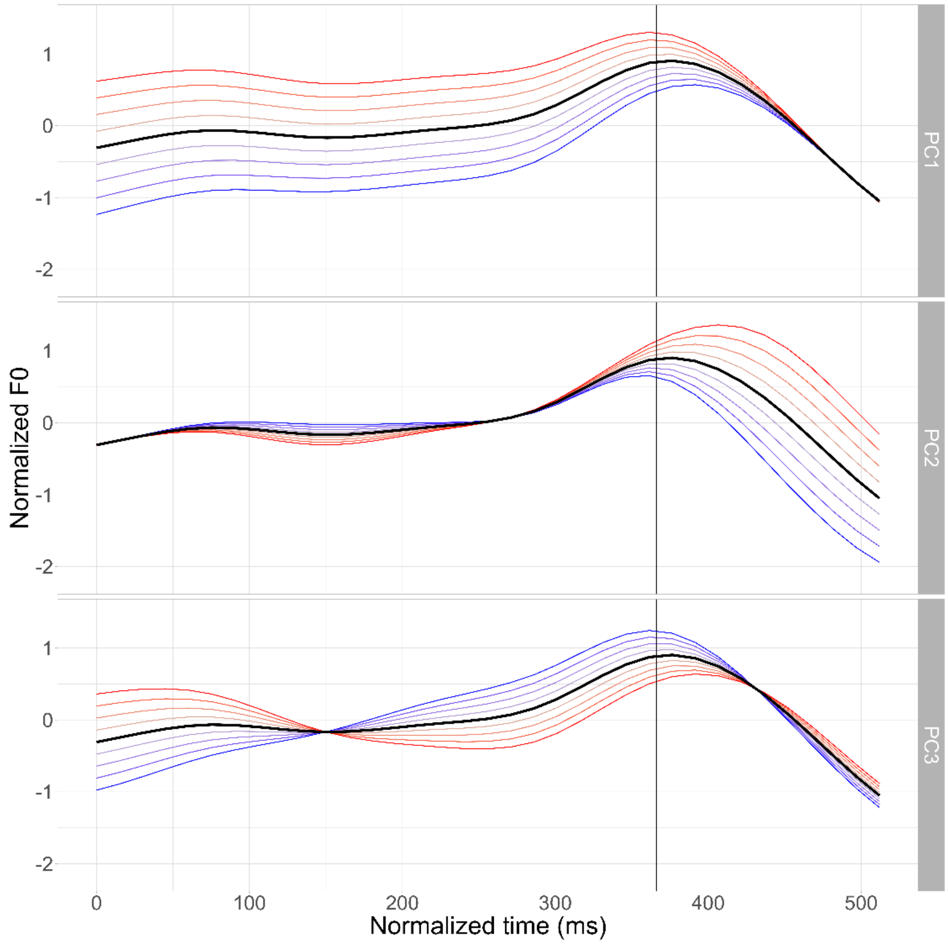


FIGURE 14. FPCA output as applied to the dataset of L+H* F0 curves. Color-coded curves illustrate the effect of each PC on the mean curve (solid black line) and were obtained by adding to the mean curve a specific PC curve multiplied by a PC score, as per Eq. (1); the equation for the curves in each panel is $f(t) = \mu(t) + \alpha \times \sigma \times (s_n) \times PC_n(t)$, where σ is the standard deviation of the PC coefficient, and α determines the fraction of the standard deviation shown. Alpha ranges from -1 (blue) to +1 (red) in increments of 0.25. The mean curve corresponds to alpha = 0 and is the same in all panels. The vertical line indicates the onset of the accented vowel.

PC	% of variance captured by PC	R^2m by PC score	R^2c by PC score
PC1	53.3%	0.32	0.32
PC2	24.6%	0.56	0.64
PC3	16.3%	0.34	0.58
PC4	3.4%	0.13	0.31
PC5	1.5%	0.09	0.27

TABLE 4. Percentage of curve variance explained by each PC resulting from the FPCA of L+H* F0 curves, followed by marginal and conditional R^2 as calculated by the LMER fitted to the scores of each PC.

We ran LMER models with the scores of the first five PCs as the dependent variable, and Stress (antepenultimate, penultimate, final) and Accentual Context (present vs. absent) as the fixed effects. Their interaction was not included as the two factors reflect pressures of different origins and at different time-points and thus are unlikely to interact. Antepenultimate was set as the reference level for Stress followed by penultimate and then final. For Accentual Context the reference level was set to *present*. Full models incorporated random intercepts for Speaker and by-Speaker random slopes for the fixed effects. Item was not included in the random structure, because it fully overlapped with Stress and Accentual Context combined. Reduction of the random structure followed the same procedure as for the models in section 3.1. For the same reasons as in the model for PC1 scores in 3.1., Speaker was removed from the random structure of the PC1 scores model here. The percentage of curve variance explained by each of the five PCs is shown in Table 4, together with R^2m and R^2c of the LMERs on their scores. Given these results, we focus on PC1, PC2, and PC3 which explained most of the curve variance and had the highest R^2 values. The LMERs for PC1, PC2 and PC3 scores are presented in Tables 5, 6a-b, and 7a-b respectively, and illustrated in Figure 15 (for FPCA output including higher PCs, and the LMERs on their scores, see https://osf.io/sjxuq/?view_only=985aa03025524b71b8a0dbc801d7a164).

Fixed Effects	Estimate	Std. Error	t-value	Pr(>|t|
(Intercept)	8.0229	1.4275	5.62	***
Stresspenult	-0.1698	1.702	-0.1	0.921
Stressfinal	1.4219	1.6945	0.839	0.402
AccentualContextabsent	-16.2509	1.3817	-11.761	***

TABLE 5. LMER results for PC1 scores from the FPCA analysis of the L+H* dataset. Final formula in R: lm(PC1 ~ Stress + AccentualContext, data = LH_1D). Key to significance levels: $p < .05$ = *; $p < .01$ = **; $p < .0001$ = ***.

Fixed Effects	Estimate	Std. Error	df	t-value	Pr(>|t|
(Intercept)	7.0624	1.048	30.071	6.734	***
Stresspenult	-7.116	0.837	285.122	-8.494	***
Stressfinal	-17.448	0.835	285.46	-20.891	***
AccentualContextabsent	2.317	0.681	285.774	3.398	***

TABLE 6A. LMER results for PC2 scores from the FPCA analysis of the L+H* dataset. Final formula in R: lmer(PC2 ~ Stress + AccentualContext+(1 | Speaker), data = LH_1D, REML=FALSE). Key to significance levels: $p < .05$ = *; $p < .01$ = **; $p < .0001$ = ***.

contrast	estimate	SE	df	t.ratio	p.value
antepenult - penult	7.12	0.842	288	8.449	<.0001
antepenult - final	17.45	0.840	289	20.778	<.0001
penult - final	10.33	0.83	288	12.455	<.0001

TABLE 6B. Results of the pairwise comparisons for PC2 scores.

Fixed Effects	Estimate	Std. Error	df	t-value	Pr(>|t|
(Intercept)	4.89	1.306	16.314	3.744	**
Stresspenult	1.097	0.732	272.219	1.498	
Stressfinal	-2.945	0.732	273.209	-4.024	***
AccentualContextabsent	-8.57	1.4	13.0762	-6.118	***

TABLE 7A. LMER results for PC3 scores from the FPCA analysis of the L+H* dataset. Final formula in R: lmer(PC3 ~ Stress + AccentualContext + (AccentualContext | Speaker), data = LH_1D, REML=FALSE). Key to significance levels: $p < .05$ = *; $p < .01$ = **; $p < .0001$ = ***.

contrast	estimate	SE	df	t.ratio	p.value
antepenult - penult	-1.10	0.735	274	-1.492	0.1368
antepenult - final	2.95	0.735	274	4.006	0.0002
penult - final	4.04	0.724	274	5.583	0.0001

TABLE 7B. Results of the pairwise comparisons for PC3 scores.

As Table 5 shows, Stress did not affect PC1 scores, but there was a significant effect of Accentual Context, such that scores were significantly higher and largely positive when a prenuclear accent preceded L+H*, but largely negative when no accent preceded (see Figure 15). The shape of the PC1 curves indicates that it was not just the F0 peak that was lower when no accent preceded but the entire F0 curve (Figure 14).

For PC2 scores, the model returned main effects of both Stress and Accentual Context (Table 6). With respect to Stress, PC2 scores significantly decreased as stress moved closer to the end of the utterance. As indicated in Figure 15, scores were mostly positive when stress was on the antepenult or the penult, while they were negative when stress was on the ultima. This difference sets curves with extreme tonal crowding apart from the rest: the accentual peak was reached earlier and scaled lower, while the following F0 drop was more precipitous (cf. Figure 14). With respect to Accentual Context, PC2 scores showed the reverse pattern to that of PC1, namely significantly higher scores when there was no preceding accent, though these differences were minimal (see Figure 15).

Finally, for PC3 scores, there were again effects of both Stress and Accentual Context (Table 7a-b and Figure 15). The Stress effect was similar to that on PC2 scores: when stress was on the final syllable, PC3 scores were largely negative, resulting in an earlier peak and a less pronounced F0 dip compared to penultimate and antepenultimate stress (Figure 14). The differences due to Accentual Context led to distinct shapes in the pre-accentual F0 stretch: when there was a preceding accent, the F0 dip was more pronounced (PC3 scores were positive) than when there was no preceding accent (PC3 scores were negative); see Figure 14.

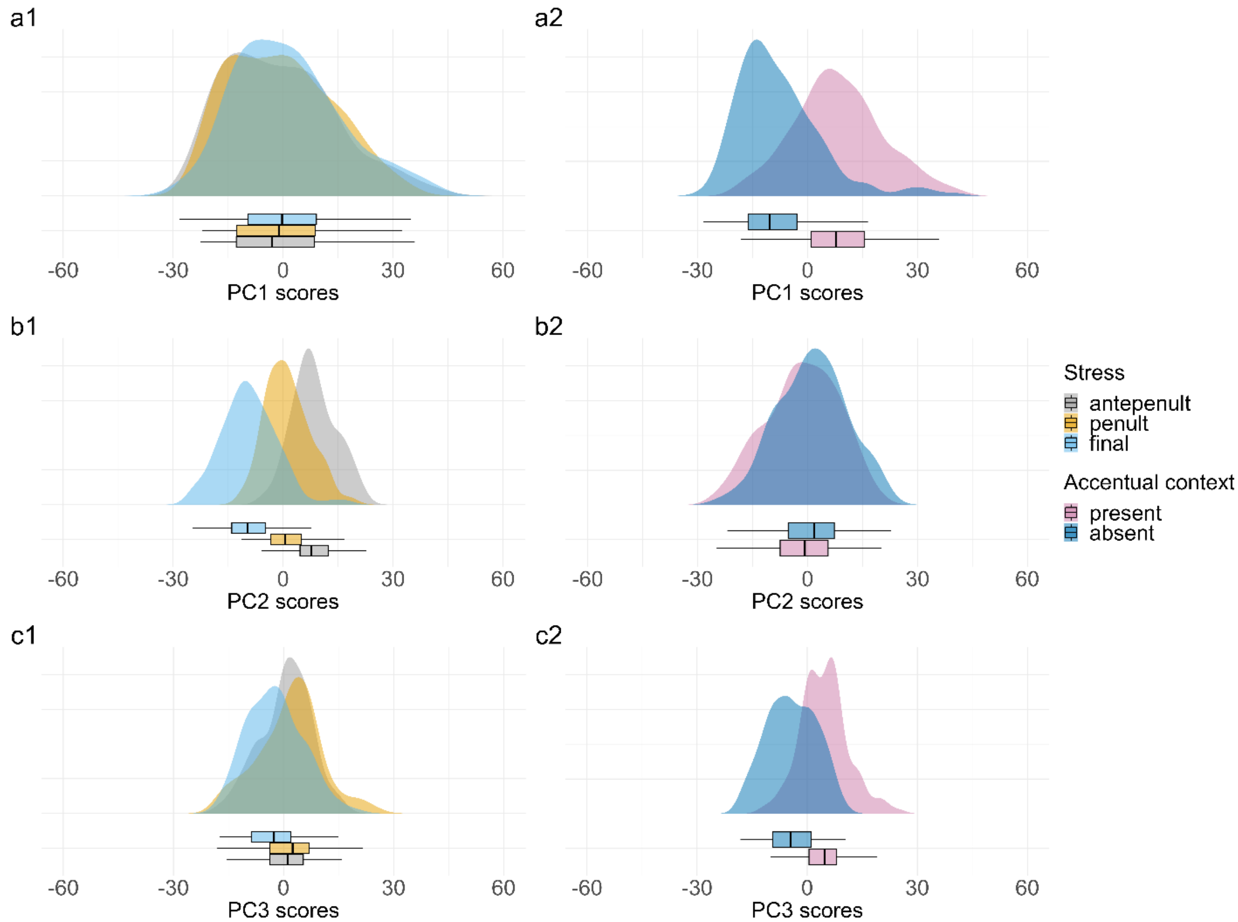


FIGURE 15. Density and box plots of PC scores for L+H* by Stress (left panels) and Accentual Context (right panels) for PC1 scores in row (a), PC2 scores in row (b), and PC3 scores in row (c).

These Stress- and Accentual context-dependent F0 shapes are reflected in the curve reconstructions in Figure 16a-b, which are based on the first three PCs and are remarkably similar to the observed averages shown in Figure 16c-d. In conclusion, the reconstruction confirms the interpretation of the FPCA output and matches the observed data.

3.4 FPCA OF L+H* CURVES INTERIM DISCUSSION: EFFECT OF STRESS AND ACCENTUAL CONTEXT

FPCA on L+H* captured not only this accent's rise-fall shape, but also fine-grained changes to its phonetic realization related to two linguistic factors, stress location and the presence of a preceding accent. The effect of Stress was captured by PC2, which indicated that earlier peaks covaried with lower scaling and a steeper F0 drop; see the curve reconstructions in Figure 16a. This combination can be attributed to tonal crowding due to the L-L% edge tones following L+H*: in the test word with final stress [ɣalaˈna], F0 is meant to drop to the bottom of each speaker's range by the end of the accented syllable (similarly to the English tune in Figures 1 and 2); in contrast, in [laðoˈlemono] the F0 drop can be delayed since there are two more syllables after the accented one on which a speaker can reach their F0 baseline (Liberman & Pierrehumbert 1984).

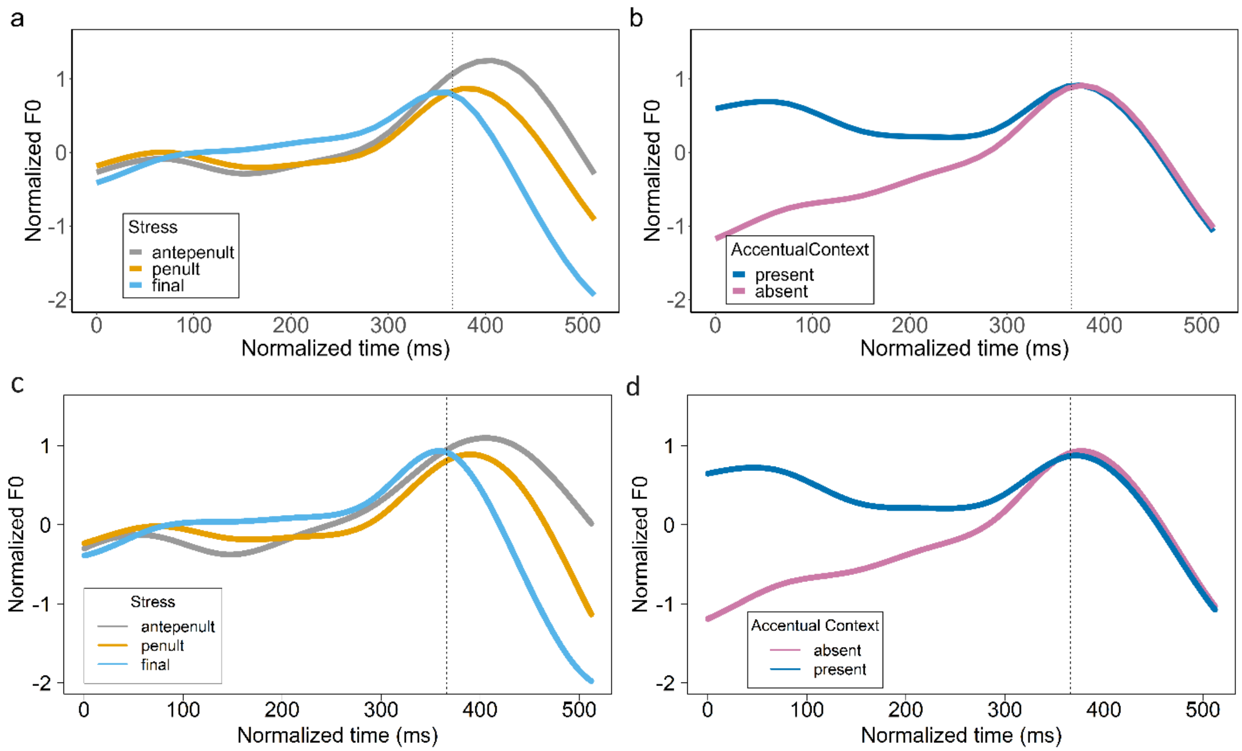


FIGURE 16. Reconstructed L+H* F0 curves by Stress in (a) and Accentual Context in (b), using PC1, PC2, and PC3 and obtained by incorporating the average value of each PC score, as predicted by the linear model, into the FPCA equation (Eq. (1)). The equations for the three curves in (a) are: for antepenult stress, $f(t) = \mu(t) - 0.10 \times PC1(t) + 8.22 \times PC2(t) + 0.60 \times PC3(t)$; for penult stress, $f(t) = \mu(t) - 0.27 \times PC1(t) + 1.10 \times PC2(t) + 1.70 \times PC3(t)$; for final stress, $f(t) = \mu(t) + 1.31 \times PC1(t) - 9.22 \times PC2(t) - 2.34 \times PC3(t)$. The equations for the two curves in (b) are: for the present condition, $f(t) = \mu(t) + 8.44 \times PC1(t) - 1.12 \times PC2(t) + 4.27 \times PC3(t)$; for the absent condition, $f(t) = \mu(t) - 7.81 \times PC1(t) + 1.19 \times PC2(t) - 4.29 \times PC3(t)$. Vertical dotted lines indicate the onset of the accented vowel. For comparison purposes, panels (c) and (d) display the smoothed, averaged and normalized curves of L+H* in the same conditions.

The effect of the Accentual Context was captured by PC1 and PC3, which showed that when L+H* was preceded by a prenuclear L*+H, the unstressed syllables before L+H* had higher scaling (higher PC1 scores), while the curve involved a drop to an F0 dip (higher PC3 scores). These effects can be plausibly attributed to the influence of the prenuclear L*+H and are reflected in the curve reconstructions in Figure 16b (see also Figure 16d for comparison): F0 remained high at the start of the analysis window following that accent's H tone; in turn, this led to a shallow but extended F0 dip just before the accented vowel onset. In contrast, in one-word utterances, F0 started from a lower point but without a pronounced localized dip.

Three main conclusions can be drawn from these results. First, they provide additional evidence that some aspects of tonal realization are DISTAL: here, the preceding accent affected the unaccented syllables preceding L+H*, ultimately creating distinct curves between utterances with one vs. two accents. Second, the effects of preceding and following context were captured by different PCs, and are thus independent of each other as we predicted; the scaling and alignment of intonation categories adjust to accommodate both. Finally, these results confirm our conclusion in section 3.2. that this accent's optimal representation is L+H*: its initial low F0 was evident even under tonal crowding (Figure 16a, c), suggesting it is an essential component of the accent (cf. Arvaniti et al. 2017 for arguments about the use of tonal crowding as a diagnostic criterion for tonehood).

3.5 JOINT F0-DURATION FPCA

In addition to FPCA on the F0 curves, we used FPCA to explore possible covariance of F0 and duration in the realization of L+H*. L+H* was a good candidate for exploring the duration-F0 link, as it is an accent known to affect duration in Greek (Arvaniti 2007, Arvaniti et al. 2006b, Katsika & Tsai 2021). Here, we wished to explore the possibility of a trade-off between duration and F0 shape that we had observed via visual inspection and annotation of the present and similar datasets: when L+H* is hypoarticulated (e.g. produced with a lower peak or a shallower dip), the accented vowel shows increased lengthening.

The joint F0-duration analysis was conducted based on the method outlined in Asano and Gubian (2018), which is briefly summarized below. For technical details, the reader is referred to Appendix A of Asano and Gubian (2018) and to https://osf.io/sjxuq/?view_only=985aa03025524b71b8a0dbc801d7a164 for our own scripts. The process makes use of the time-warped curves – ($h(t)$ in Eq. (2)) – which are produced by landmark registration and map normalized time (t) to original time. Each $h(t)$ is transformed into $r(t)$, using Eq. (2); $r(t)$ represents logarithmically the rate of change of the time-warped curve $h(t)$ with respect to time.

Eq. (2) $r(t) = -log(dh(t)/dt)$

The pairs of F0 and $r(t)$ curves are used as inputs to the joint FPCA (as opposed to just the F0 curves that served as the input to the FPCA in 3.3). The outputs of this analysis are expressed as shown in Eq. (3a) for F0, which is essentially the same as Eq. (1), and in Eq. (3b) for speaking rate. Both equations share the same PC scores (s1, s2, …), thus capturing co-variation between F0 and duration.

Eq. (3a) $F0(t) = \mu_{F0}(t) + s1 \times PC1_{F0}(t) + s2 \times PC2_{F0}(t) + \ldots$

Eq. (3b) $r(t) = \mu_r(t) + s1 \times PC1_r(t) + s2 \times PC2_r(t) + \ldots$

The first five PCs of this analysis explained 97.7% of the combined variance of duration and F0 curves (Table 8). We also ran LMER models using the same procedures as before and present R^2m and R^2c in Table 8. In this analysis, we were interested in the effect of Stress and Accentual Context on the relationship between duration and F0. Thus, we focus on PC1 and PC2, the only PCs that revealed a systematic relationship between F0 and duration and also had high R^2m (see Tables 9a-b and 10a-b, and Figures 17 and 18). In contrast, PC3 had low R^2m and high R^2c, suggesting it explained mostly individual variability. (For FPCA output including higher PCs, and LMERs on their scores, see https://osf.io/sjxuq/?view_only=985aa03025524b71b8a0dbc801d7a164)

PC	% of variance captured by PC	R^2m by PC score	R^2c by PC score
PC1	37.8%	0.63	0.80
PC2	31.8%	0.33	0.48
PC3	13.4%	0.03	0.48
PC4	8.1%	0.31	0.54
PC5	6.6%	0.23	0.49

TABLE 8. Percentage of curve variance explained by each PC resulting from the FPCA on the combined F0 and duration of L+H*, followed by marginal and conditional R^2 as calculated by the LMER fitted to the scores of each PC.

Based on the FPCA output, PC1 indicates that F0 curves with a less pronounced dip and an earlier and lower peak were associated with longer accented vowel duration (Figure 17). The LMER results suggest that the effect was due to stress and applied when stress was on the final or the penultimate syllable, so that curves associated with these stress positions had negative PC1 scores (see Table 9a-b and Figure 18). Conversely, when stress was on the antepenult, PC1 scores were strongly positive: peaks were later and higher and associated with shorter accented vowels.

Further, PC2 shows that lower F0 curves, particularly those low in the preaccentual region, were associated with longer duration of that region (Figure 17). The LMER output indicates that the differences were due to Accentual Context: PC2 scores were largely positive when context was present but largely negative when it was absent (see Table 10a-b and Figure 18), indicating that lower F0 scaling was associated with longer duration in one-word utterances and that the reverse obtained in two-word utterances.

The above-mentioned trade-off between F0 and duration were supported by the reconstructions in Figure 19. These show that lower scaling due to tonal crowding was associated with longer accented vowel duration (PC1 in panel a), and that when additional material led to shorter durations, F0 increased (PC2, panel b).

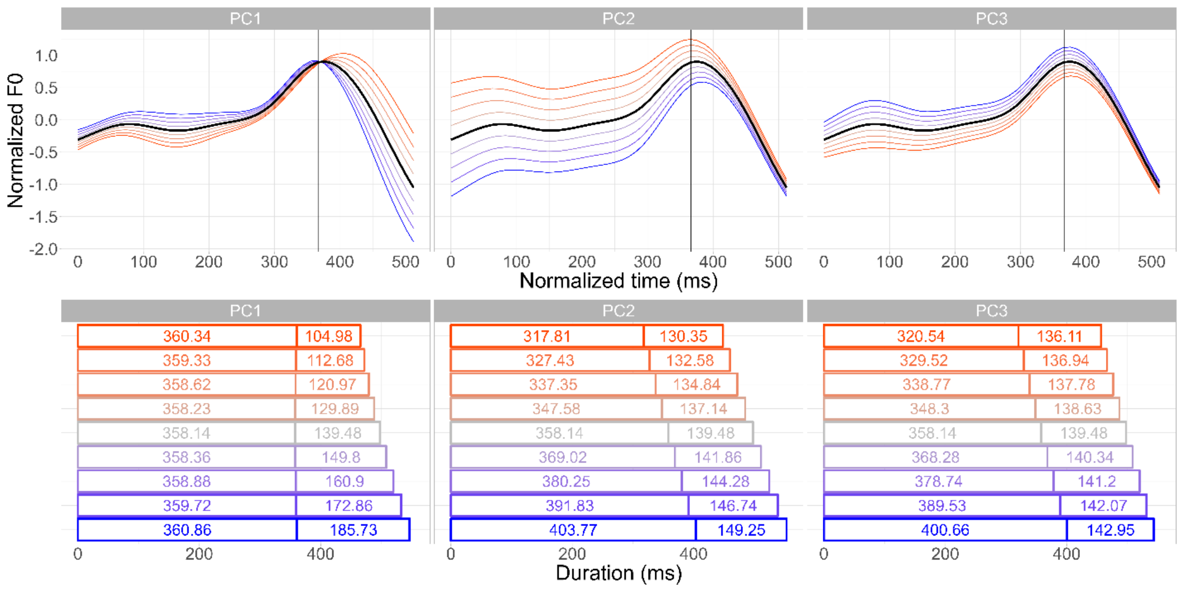


FIGURE 17. Joint F0-duration FPCA output as applied to the dataset of L+H* F0 curves (top) and corresponding durations (bottom). Top panel: Color-coded curves illustrating the effect of each $\mathrm{PC_{F0}}$ on the mean F0 curve (solid black line), and obtained by adding to the mean curve a specific $\mathrm{PC_{F0}}$ curve multiplied by a $\mathrm{PC_{F0}}$ score, as shown in Eq. (3a); the equation for the $\mathrm{F0(t)}$ curves in each panel is $\mathrm{F0(t)} = \mu_{\mathrm{F0}}(\mathrm{t}) + \alpha \times \sigma \times (\mathrm{s_n}) \times \mathrm{PC_{F0n}}(\mathrm{t})$, , as shown in Eq. (3a); σ is the standard deviation of the PC coefficient, and α determines the fraction of the standard deviation, ranging from -1 (blue) to +1 (red) in increments of 0.25. The mean curve corresponds to alpha = 0 and is the same in all panels. The vertical line indicates the onset of the accented vowel. Bottom panel: color-coded bars illustrating the effect of each $\mathrm{PC_r}$ on mean duration (solid grey bar). Each bar's value was obtained by solving for interval duration using Eq. (2), where $\mathrm{r(t)} = \mu_{\mathrm{r}}(\mathrm{t}) + \alpha \times \sigma \times (\mathrm{s_n}) \times \mathrm{PC_{rn}}(\mathrm{t})$, as shown in Eq. (3b). The PC score $(\mathrm{s_n})$ in Eqs. (3a) and (3b) are the same, as predicted by the linear model. The color-coding demonstrates varying PC scores, calculated by multiplying the PC score by fractions (α) of its standard deviation (σ), with α ranging from -1 to 1 in increments of 0.25.

Fixed Effects	Estimate	Std. Error	df	t-value	Pr(>|t|
(Intercept)	16.0297	2.3844	13.6721	6.723	***
Stresspenult	-22.8249	1.4445	16.9436	-15.802	***
Stressfinal	-31.9864	3.1661	13.0549	-10.103	***
AccentualContextabsent	4.4969	0.9045	272.85	4.972	***

TABLE 9A. LMER results for PC1 scores from the joint FPCA analysis of F0 and duration in the L+H* dataset. Final formula in R: lmer(PC1 ~ Stress + AccentualContext + (Stress | Speaker), data = LH_1D, REML=FALSE). Key to significance levels: $p < .05$ = *; $p < .01$ = **; $p < .0001$ = ***.

contrast	estimate	SE	df	t.ratio	p.value
antepenult - penult	22.82	1.49	13.3	15.33	<.0001
antepenult - final	31.99	3.30	14.1	9.706	<.0001
penult - final	9.16	2.65	13.9	3.455	=.0039

TABLE 9B. Results of the pairwise comparisons for PC1 scores.

Fixed Effects	Estimate	Std. Error	df	t-value	Pr(>|t|
(Intercept)	12.097	2.087	21.737	5.796	***
Stresspenult	-2.740	1.635	272.947	-1.676	n.s.
Stressfinal	-5.545	1.633	274.560	-3.396	***
AccentualContextabsent	-17.817	2.851	13.214	-6.250	***

TABLE 10A. LMER results for PC2 scores from the joint FPCA analysis of F0 and duration in the L+H* dataset. Final formula in R: lmer(PC2 ~ Stress + AccentualContext + (AccentualContext \| Speaker), data = LH_1D, REML=FALSE). Key to significance levels: $p < .05$ = *; $p < .01$ = **; $p < .0001$ = ***.

contrast	estimate	SE	df	t.ratio	p.value
antepenult - penult	2.74	1.64	275	1.669	n.s.
antepenult - final	5.55	1.64	276	3.379	0.0025
penult - final	2.80	1.62	275	1.735	n.s.

TABLE 10B. Results of the pairwise comparisons for PC2 scores.

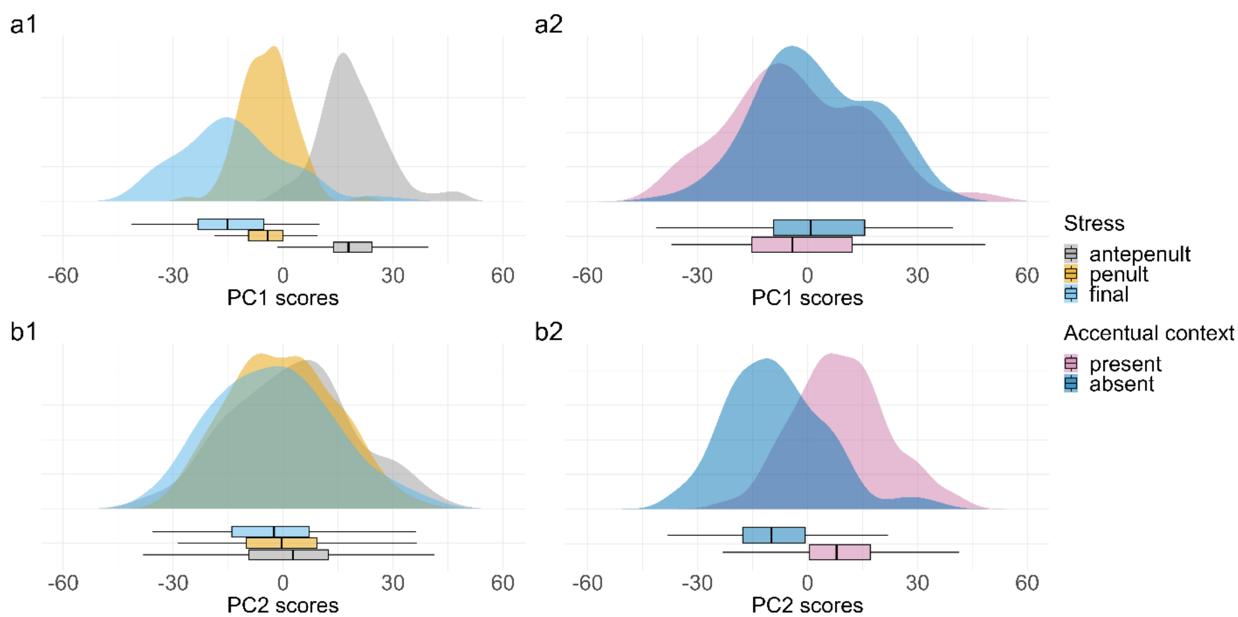


FIGURE 18. Density and box plots of PC scores from the joint F0-duration FPCA on the L+H* dataset, separately for Stress (left panels) and Accentual Context (right panels) for PC1 (row a), and PC2 (row b).

3.6 JOINT F0-DURATION FPCA INTERIM DISCUSSION: WITHIN CATEGORY TRADE-OFFS

The results of the joint F0-duration FPCA provided evidence of trade-offs between these two parameters. These are evident in Figure 17 and the reconstructions in Figure 19 and statistically confirmed. Thus, the results supported our hypothesis that hypoarticulating L+H* is compensated for by elongating the accented vowel. Additionally, we found that shorter segment duration was associated with higher peaks. This effect is particularly noteworthy because it applied to the longer utterances (when Accentual Context was present), which would be expected to show lower accentual peaks due to declination. The effect of both trade-offs is that L+H* sounds high in pitch (cf. Barnes et al. 2021).

Importantly, these trade-offs pertain to within-category variability. They are not meant to boost cross-category distinctions, like the enhancement effects mentioned in section 3.2, and similar effects reported by others, such as Roessig and Mücke (2019). Taken together, the results reported in sections 3.1 and 3.3 support the idea that enhancement is used to differentiate categories, while cue-trading helps optimize the realization of a given category under multiple constraints.

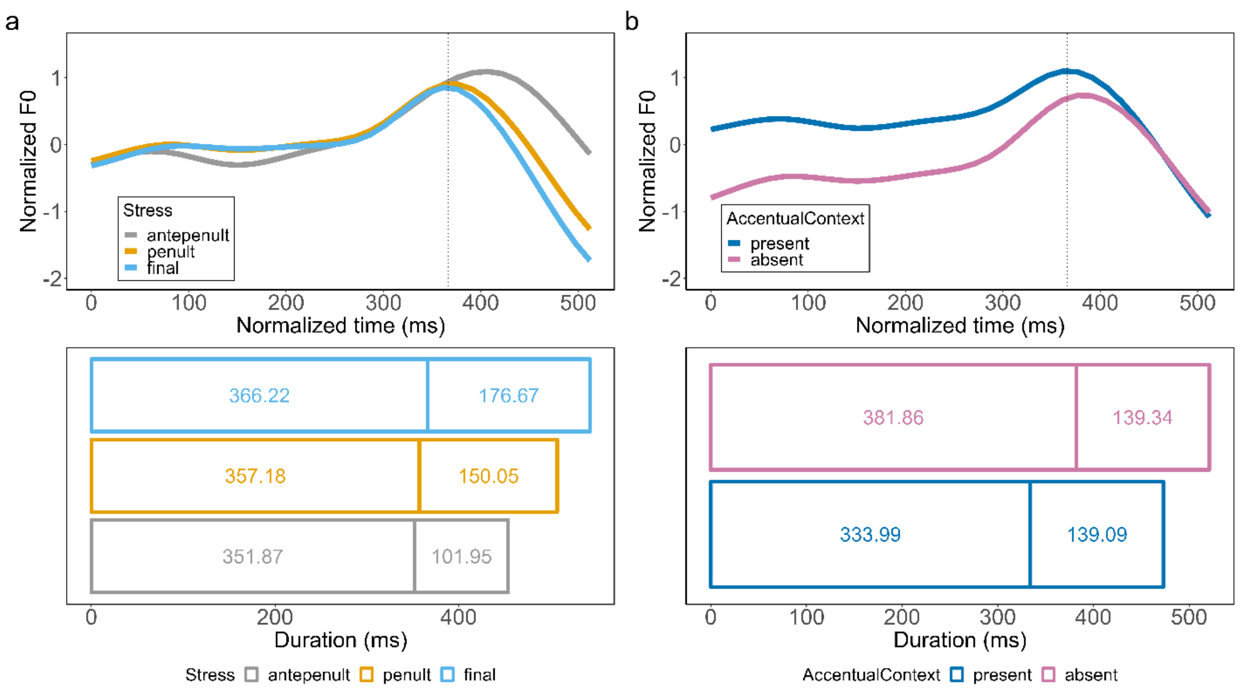


FIGURE 19. Reconstructed L+H* F0 curves (top) and corresponding durations (bottom) by stress location in (a) and accentual context in (b), using PC1 and PC2. The F0 curves were obtained by incorporating the average value of each PC score, as predicted by the linear model, into the FPCA equation (Eq. (3a)). The equations for the three curves in (a) are: for antepenult stress, $f(t) = \mu(t) + 18.27 \times PC1(t) + 3.18 \times PC2(t)$; for penult stress, $f(t) = \mu(t) - 4.54 \times PC1(t) + 0.44 \times PC2(t)$; and for final stress, $f(t) = \mu(t) - 13.70 \times PC1(t) - 2.35 \times PC2(t)$. The equations for the two curves in (b) are: for the present condition, $f(t) = \mu(t) - 2.24 \times PC1(t) + 9.33 \times PC2(t)$; for the absent condition, $f(t) = \mu(t) + 2.25 \times PC1(t) - 8.48 \times PC2(t)$. The reconstructed durations (ms) were calculated by first obtaining the $r(t)$ curves corresponding to the F0 curves listed above (i.e. with the same PC scores) using the equation in Eq. (3b). Then, by substituting the obtained $r(t)$ into Eq. (2), the duration of the intervals resulting from landmark registration positions was determined and plotted in the bottom panel.

4 DISCUSSION

Our analyses lead to a number of conclusions regarding intonation phonetics and provide a means by which variability, long considered a stumbling block in understanding and analyzing intonation, can be handled. We first discuss the relevance of our findings for understanding intonation categories and their representations, and then focus on within-category effects, before addressing the four main questions we raised in the Introduction. We conclude with recommendations based on the repercussions of our findings for intonational phonology and analysis.

Our results from the F0 FPCA on the entire dataset showed that intonation categories differ in expected dimensions, predicted by AM: scaling and F0 shape (as reflected in AM representations), and the synchronization of tonal events with the segmental string. Importantly, these features are not invariant, as often implicitly assumed, even when their role is to distinguish phonological categories, such as the three accents in the present study. As our results show, there is significant overlap in phonetic values, though this overlap is not more extensive than that reported for vowel contrasts (cf. Strange et al. 2007, Freeman 2021). This similarity is notable because intonation has long been considered qualitatively different from segments and often seen as extremely variable. Our findings argue against this position: intonation is no more variable than other speech categories, once appropriate tools are employed for its investigation (cf. Fox & Jacewicz 2009 on the usefulness of dynamic measures in the investigation of vowel contrasts).

In addition, our results show that the realization of intonation categories does not rely exclusively on local F0 targets, as often assumed in AM phonetics. In the first instance, realization may involve distal F0 cues that enhance main tonal events. In the present data, H*+L differed from H* and L+H* in terms of the high scaling of the syllables preceding the accent. Since the main difference between H* and H*+L is scaling, and plateaux sound higher than peaks (Knight 2008), this distal difference functions as a type of enhancement. Similarly, the FPCA of the L+H* F0 curves showed a difference between curves preceded by another accent and those that were not, such that the former showed an extended F0 depression (Figure 16b, d). A way to understand this feature is as a means of enhancing the L tone by adding to its duration. It remains to be seen if such distal F0 cues are used by listeners to predict upcoming tonal events but their presence in production is clear.

Second, our results show evidence of cue trading between F0 and duration, a relationship not previously reported for intonation categories: the undershooting of L+H* was compensated for by the elongation of the accented vowel, while the faster speaking rate of longer utterances, which would lead to relatively short accented vowels, was countered with higher F0 scaling. Evidence of cue enhancement and cue trading should not be surprising: Cue redundancy characterizes all aspects of speech, and cognition more generally (e.g. Parker, Diehl & Kluender 1986, McMurray & Jongman 2011, Martin 2016). The use of cues beyond F0 is increasingly documented in the study of lexical tones which are enhanced by means of phonation, duration, and intensity (on phonation, Zheng 2006 on Mandarin, Khan, Xu & Sohail 2020 on Pahari, Wang, Gussenhoven & Liang 2020 on Kaifeng Mandarin, Zhang & Kirby 2020 on Cantonese; on duration, Khan et al. 2020 on Pahari, Yu 2023 on Cantonese, Wu, Adda-Decker & Lamel 2023 on Mandarin; on intensity, Khan et al. 2020 on Pahari). Such additional cues are, nevertheless, only beginning to be considered in intonation research which has so far focused almost exclusively on F0. In this respect, our work adds to calls to avoid an 'over-reliance on the F0 trajectory' (Albert et al. 2018: 804), and examine the integration of F0 and non-tonal cues, such as duration (Steffman & Jun 2019) and periodic energy (Albert et al. 2018), while paying greater attention to the perceptual effects of F0 cues (Barnes et al. 2021). Based on this body of work and our own findings, intonation emerges as a phonetically multi-dimensional system that exhibits cue redundancy, cue enhancement, and cue trading. In turn, this means that intonational categories are encoded by a rich system of cues, much like segmental and lexical tonal contrasts, making between-category overlap of specific melodic dimensions of less importance than sometimes thought.

Within-category, the investigation of L+H* showed evidence of scaling, alignment and shape variation under the influence of two linguistic factors, tonal crowding and accentual context. Like previous studies (e.g. Arvaniti & Ladd 2009), we found that such effects are systematic; yet, they still contribute to within-category variability and by extension to cross-category overlap and thus they need to be addressed when studying intonation. Further, not all variability is linguistically driven. As an illustration, in the analysis of L+H* F0 curves, 34% of PC3 score variance was due to Stress and Accentual Context (R^2m in Table), another 24% was explained by speaker variability (R^2c-R^2m), the only random factor in the LMER, while the remainder was probably due to noise (or, potentially, sources not covered in our modelling).

These conclusions would not be easy to reach without the use of FPCA, which offers distinct advantages as a tool for investigating intonation. Notably, the separation of curve analysis, which is data driven, from statistical modelling, which relies on linguistic categories, made it possible to separate general variability from linguistically-driven variation (Bürki 2018), and thus avoid circularity in accent categorization, at least with the methodology used here, in which the linguistic classification was based on intended pragmatic meaning, not accent shape. Further, by determining the size of the analysis window and the landmark around which F0 curves were time-normalized, we could focus on curves that fitted AM's principles and our own research questions. In particular, choosing the onset of the accented vowel allowed us to incorporate the fundamental AM insight that tonal events align with segmental landmarks, while selecting a larger analysis window than is typical in AM allowed us to document distal and dynamic F0 changes. Our final analysis allowed us to investigate cue trading by examining F0 and duration covariance. Though these choices reflect our own research program, FPCA is sufficiently flexible to fit different research questions and frameworks. Most importantly, variability was rendered manageable by using FPCA, when it would have seemed insurmountable if examined based on static, highly localized targets. Overall, our study adds to an increasing body of work showing that dynamic, data-driven methods of curve analysis that rely on dimension-reduction are a promising way forward for intonation research (Grabe et al. 2007, Chen and Boves 2018, Asano and Gubian 2018, Baltazani et al. 2022). Such methods may also provide a breakthrough in understanding languages with highly variable realization of intonation categories, such as

Serbian and Croatian (Smiljanic 2006), Dalabon (Fletcher 2015), Mawng (Fletcher et al. 2016), Ambonese Malay (Maskikit-Essed & Gussenhoven 2016) and Tashlhiyt Berber (Roettger 2017).[7] They may also be particularly helpful in modelling lexical tone and pitch accent systems, tone-to-tone coarticulation, and the challenging relationship between tone and intonation (e.g. FPCA can be used to test if F0 changes attributed to tone vs. intonation are independent of each other).

Although FPCA is data-driven, a critical element of the present study is that it was employed in conjunction with a conceptualization of how to view and investigate intonation. This approach consists in a set of a principles, stemming from a main postulation condensed into *Tame INTonation* (TINT): intonation is structured in a way comparable to that of other subsystems of phonology, not an 'untamed savage' (Bolinger 1978: 475) that is found 'around the edge of language' (Bolinger 1964: 282). This follows from the central AM postulate that intonation is part of a language's phonology. A number of corollaries stem from this main point.

First, in order to develop phonological representations of intonation, it is essential to abstract away from phonetic detail so as to capture what is contrastive in a given intonational system, not document, formalize and discretize context-dependent variability. In order to reach this analytical outcome, the investigation of intonational phonetics requires the use of methods that reduce variability, such as FPCA, provided they are not used exclusively on pre-determined phonetic categories (or the results will be circular). Thus, for this type of analysis to work, it is necessary to enlist pragmatics (e.g. Pierrehumbert & Hirschberg 1990, Steedman 2000), as was done here and in other recent work (e.g. Gryllia et al. 2018, Baltazani, Grylllia & Arvaniti 2020, Kim et al. 2023). Recall that in the present study, we categorized the accents based on the pragmatic context in which they were elicited; doing so avoided the circularity that would result from categorizing the accents based on their phonetics and then seeking to validate our choices based on the accents' phonetic features.

Most importantly, fully accepting that intonation is part of phonology means that the way to approach intonation should not be fundamentally different from that employed for segments. As our study shows, and as the body of work reviewed here also attests, intonational categories do not fundamentally differ from vowels and consonants. Consequently, parallelisms with linguistic approaches to segmental structure should be given due consideration when studying intonation phenomena, and the study of intonation should rely on criteria and assumptions similar to those used for the study of segments. For instance, dialectal differences should be considered just as they are when segments are investigated. Similarly, the practice of relying on just one phonetic dimension is not typically adopted in the study of segmental contrasts, and there is no reason why it should be adopted in the study of intonation. Sources of variability

[7] Following Pierrehumbert & Beckman (1988), Arvaniti (2022b) argues that such differences may give rise to a typological distinction between languages in which tonal association percolates from metrical positions (such as heads and phrasal boundaries) to specific TBUs, leading to segmental anchoring, and those in which such percolation does not take place, giving rise to loose alignment. Independently of whether this typological distinction holds, accounting for cross-linguistic differences between stable and variable realization of tunes is imperative for understanding intonation.

should be studied, modelled and understood (as they are with respect to segments), but they need not be the sole focus of inquiry or a point of concern. At the same time, particularities of intonation should be taken into consideration: There is now mounting evidence that the processing of intonation is probabilistic and relies on context (Calhoun 2010; see Kurumada & Roettger 2022 for a review). This should be taken into account when intonation production is considered and perception experiments are planned.

The above principles and conclusions from our study address the questions we posited in the introduction. First, they argue that AM is correct in separating phonetics from the phonology of intonation. This allows AM to capture significant generalizations by providing both abstract representations and a blueprint for phonetic realization (question A). Further, by using the methodology presented here it is possible to document and statistically model linguistically-driven variation and distinguish it from general phonetic variability and noise (question B). These two points amount to an overall system that can be learnt by a language's speakers and account for their intuitions (question C): it is a system that relies on abstract tunes associated with metrical structure, on the one hand, coupled with rich phonetics on the other, and processed probabilistically. Such a conceptualization points to variability being less of a problem than it is often considered to be when phonetics and phonology are conflated. This conceptualization is in line with the current understanding that variability is not only unproblematic but essential for learning and generalization (e.g. Bürki 2018, Raviv, Lupyan & Green 2022). Thus, it is high time we distinguished between cognitive requirements for a linguistic system from difficulties faced by researchers when analyzing phonetic data: for the former, variability is required, even if for the latter it is a considerable obstacle. The method presented here shows how the two can be reconciled.

The above still leaves us with question D: how can we make a principled distinction between gradience, noise and categorial differences? Our study did not explicitly address gradience, in that gradience as an expression of paralinguistic distinctions was not part of the study design. However, if we adhere to the principles and practices presented here, gradience ceases to be a major issue. Unsystematic variability, in other words noise that is unlikely to be purposeful, can be filtered out: this is variability explained by a PC but shown to be unrelated to linguistic predictors. Systematic aspects of gradient, paralinguistic uses of F0 – such as raising pitch or delaying peaks for politeness (Borràs-Comes, Sichel-Bazin & Prieto 2015, Gryllia et al. 2018) – can be modelled statistically. For example, tunes can be investigated in communicative situations that require different degrees of politeness and statistically modelled to explore politeness effects on PC scores. Given their origin, however, such effects should not be confused with intonation per se: they pertain to the use of F0, not intonation structure, and thus analyses should not attempt to include them into a language's intonational grammar, much as we do not include into phonological representations the effects of, say, smiling on segments (cf. Hayes 1994, Gussenhoven 2004, chap. 4, Ladd 2014, chap. 4 and references therein). Thus, in Greek (as in English), if one wishes to insist on a correction, they may raise the F0 of a L+H* accent, delay its peak, do both, speak louder or more slowly, or do all of the above; all these options are adopted purposefully to indicate the speaker's insistence. This, however, does not mean that we need to posit multiple accents to account for these effects: we would not posit extra loud /e/ as a phonological category and the same should apply to intonation.

To conclude, gradience should be understood as distinct from structural aspects of intonation, based on a principled separation of attitudinal functions of F0 from intonational meaning proper (treated formally, as in Pierrehumbert & Hirschberg 1990, or Steedman 2000). Based on the above, we can return to our original question about whether the black F0 curve of *an anemone* in Figure 2b is an instance of a distinct tune. Our answer would be no: this curve was produced as a response to a question requesting new information and it was simply produced with a greater pitch span than the rest, a change that also resulted in some peak delay. It is possible that listeners would detect greater assertiveness in this token relative to the others, but they would interpret it in context, and thus treat it as phonologically equivalent to them. Similarly, although we confirmed a scaling difference between H* and H*+L, this applied to the corpus, not to individual tokens: some tokens, depending on the speaker's voice and their degree of excitement or surprise etc., will be higher than others and this will inevitably lead to within-category gradience and potential cross-category overlap. We argue, however, that this should be neither a point of concern nor the focus of intonation analysis.

Inevitably, gray areas will always be present. For instance, some uses of F0 may be in the process of being grammaticalized, and thus they may be neither phonological nor strictly a matter of phonetic realization, thereby exhibiting gradience that is both systematic and difficult to unequivocally classify as phonetic or phonological (see Gussenhoven & van de Ven 2020 on Zhumadian Mandarin). We contend, however, that such cases will be the exception, not the norm, once a stricter approach, like that advocated here, is adopted for positing phonological categories of intonation. At any rate, such cases are not unique to intonation, and thus they should not be granted undue attention (see Scobbie 2007 for a similar discussion with respect to segmental contrasts).

5 CONCLUSION

In conclusion, this body of work showed that the realization of intonation is characterized by sizeable category overlap, even when examined by means of a dimension-reduction method like FPCA. Nevertheless, tonal categories remain distinct thanks to cue redundancy and cue-trading. Further, intonation categories evince both systematic changes, driven by linguistic factors such as tonal context and crowding, and other types of variability. These results have repercussions for standard practices and assumptions in the study of intonation. They suggest that we cannot rely on one dimension, such as peak alignment, to distinguish tonal events, as values in any one dimension cannot be uniquely attributed to a single tonal category. Instead, we need to engage in a richer investigation of intonation and consider parameters beyond F0.

The picture that emerges from the above is that tonal events are comparable to segments: they are realized by means of several phonetic dimensions that can be in enhancing or trading relationships with each other and exhibit within-category variability and cross-category overlap. This variability should not be seen as a concern. First, variability is essential for learning and processing categories; intonation cannot possibly be an exception to such a basic cognitive requirement. In turn this means that variability is not an issue for learners and language users, but only for analysts. In this respect, variability ceases to be a problem if we adopt a principled separation of linguistically driven variation and other sources of variability, stop relying on unobtainable tonal target invariance for analyses, and adopt suitable analytical and statistical tools. Here we have shown that this can be done if we adopt a method like FPCA and split the analysis of general curve variability from statistical modelling based on linguistic categories. This allows us to partition variability into linguistically relevant (or otherwise systematic) variation, on the one hand, and residual variability, such as noise, and speaker- and item-specific differences, on the other. Furthermore, this separation allows for the investigation of residual variability, such as gradient paralinguistic effects, should these be the research focus.

Adopting a methodology like the one presented here will be maximally fruitful if accompanied by three additional stipulations: the need to meticulously differentiate intonation as a component of phonology from F0, its main (but not its sole) phonetic exponent; the need to distinguish between streamlined abstract phonological representations of intonation and their phonetic realization, which involves not just F0 but also non-tonal parameters, such as duration; the recognition that, in this conceptualization, intonational meaning is critical for determining intonational categories in a non-circular fashion. In sum, what we advocate for is an approach to the study of intonation that adopts the same assumptions, principles, and practices we employ in the study of segmental contrasts, rather than treating intonation as a 'half-tamed savage' that is peripheral to the study of speech.

References

Ahn, Byron; Nanette Veilleux; Stefanie Shattuck-Hufnagel; and Alejna Brugos. Submitted 2021. *Embarking on PoLaR Explorations: A Framework for Intonational Annotation and Analysis.* Online: https://osf.io/cj4av/.

Albert, Aviad; Francesco Cangemi; and Martine Grice. 2018. Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. *Proceedings of the 9th International Conference on Speech Prosody (Speech Prosody 2018)*, Poznań, 804–808. Online: https://www.isca-speech.org/archive/pdfs/speechprosody_2018/albert18_speechprosody.pdf.

Arvaniti, Amalia. 2007. On the relationship between phonology and phonetics (or why phonetics is not phonology). *Proceedings of the16th International Congress of Phonetic Sciences (ICPhS 2007)*, Saarbrücken, 19–24.

Arvaniti, Amalia. 2016. Analytical decisions in intonation research and the role of representations: Lessons from Romani. *Laboratory Phonology* 7(2).6. Online: http://dx.doi.org/10.5334/labphon.14.

Arvaniti, Amalia 2022a. The Autosegmental-Metrical model of intonational phonology. In Barnes & Shattuck-Hufnagel, 25–63.

Arvaniti, Amalia 2022b. Author response to the commentary: In Barnes & Shattuck-Hufnagel, 76–83.

Arvaniti Amalia, and Mary Baltazani. 2005. Intonational analysis and prosodic annotation of Greek spoken corpora. In Jun, 84–117.

ARVANITI, AMALIA, and GINA GARDING. 2007. Dialectal variation in the rising accents of American English. *Papers in Laboratory Phonology 9*, ed. by Jennifer Cole and José Ignacio Hualde, 547–576. Berlin: Mouton de Gruyter.

ARVANITI, AMALIA, and D. ROBERT LADD. 2009. Greek wh-questions and the phonology of intonation. *Phonology* 26.43–74.

ARVANITI, AMALIA, and D. ROBERT LADD. 2015. Underspecification in intonation revisited: a reply to Xu, Lee, Prom-On and Liu. *Phonology* 32.537–541.

ARVANITI, AMALIA; D. ROBERT LADD; and INEKE MENNEN. 1998. Stability of tonal alignment: The case of Greek prenuclear accents. *Journal of Phonetics* 26.3–25.

ARVANITI, AMALIA; D. ROBERT LADD; and INEKE MENNEN. 2000. What is a starred tone? Evidence from Greek. *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, ed. by Michael Broe and Janet Pierrehumbert, 119–131. Cambridge: Cambridge University Press.

ARVANITI, AMALIA; D. ROBERT LADD; and INEKE MENNEN. 2006a. Phonetic effects of focus and "tonal crowding" in intonation: Evidence from Greek polar questions. *Speech Communication* 48.667–696.

ARVANITI, AMALIA; D. ROBERT LADD; and INEKE MENNEN. 2006b. Tonal association and tonal alignment: Evidence from Greek polar questions and contrastive statements. *Language and Speech* 49.421–450.

ARVANITI, AMALIA; MARZENA ŻYGIS; and MAREK JASKUŁA. 2017. The phonetics and phonology of the Polish calling melodies. *Phonetica* 73.338–361.

ASANO, YUKI, and MICHELE GUBIAN. 2018."Excuse meeee!!": (Mis)coordination of lexical and paralinguistic prosody in L2 hyperarticulation. *Speech Communication* 99.183–200.

ATTERER, MICHAELA, and D. ROBERT LADD. 2004. On the phonetics and phonology of "segmental anchoring" of *F0*: Evidence from German. *Journal of Phonetics* 32.177–197.

BALTAZANI, MARY; STELLA GRYLLIA; and AMALIA ARVANITI. 2020. The intonation and pragmatics of Greek wh-questions. *Language and Speech* 63.56–94.

BALTAZANI, MARY; JOHN COLEMAN; ELISA PASSONI; and JOANNA PRZEDLACKA. 2022. Echoes of past contact: Venetian influence on Cretan Greek intonation. *Language and Speech*, 0(0). Online: https://doi.org/10.1177/00238309221091939

BARNES, JONATHAN; NANETTE VEILLEUX; ALEJNA BRUGOS; and STEFANIE SHATTUCK-HUFNAGEL. 2010. Turning points, tonal targets, and the English L- phrase accent. *Language and Cognitive Processes 25*.982–1023.

BARNES, JONATHAN; ALEJNA BRUGOS; NANETTE VEILLEUX; and STEFANIE SHATTUCK-HUFNAGEL. 2021. On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics* 85.101020. Online: https://doi.org/10.1016/j.wocn.2020.101020.

BARNES, JONATHAN, and STEFANIE SHATTUCK-HUFNAGEL (eds.) Prosodic theory and practice. Cambridge, MA: The MIT Press.

BARTOŃ, KAMIL. 2022. MuMIn: Multi-Model Inference. R package version 1.46.0. Online: https://CRAN.R-project.org/package=MuMIn

BATES, DOUGLAS; MARTIN MAECHLER; BEN BOLKER; and STEVE WALKER. 2015. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software* 67.1–48.

BECKMAN, MARY E., and GAYLE AYERS ELAM. 1997. *Guidelines for ToBI Labelling* (v.3). The Ohio State University Research Foundation.

BECKMAN, MARY E., and JANET B. PIERREHUMBERT. 1986. Intonational structure in Japanese and English. *Phonology* 3.255–310.

BECKMAN, MARY E., and JENNIFER J. VENDITTI. 2010. Tone and intonation. In Hardcastle et al., 603–632.

BOERSMA, PAUL, and DAVID WEENINK. 2018. *Praat: doing phonetics by computer* [Computer program].

BOLINGER, DWIGHT L. 1964. Around the edge of language: intonation. *Harvard Educational Review* 34(2).282–296.

BOLINGER, DWIGHT L. 1978. Intonation across languages. *Universals of Human Language* (vol.2), ed. by Joseph Greenberg, 471–524. Stanford: Stanford University Press.

BOLINGER, DWIGHT L. 1986. *Intonation and its parts: Melody in spoken English*. London: Arnold.

BORRÀS-COMES, JOAN; RAFÈU SICHEL-BAZIN; and PILAR PRIETO. 2015. Vocative intonation preferences are sensitive to politeness factors. *Speech Communication* 58.68–83.

BROWMAN, CATHERINE P., and LOUIS GOLDSTEIN. 1992. Articulatory phonology: An overview. *Phonetica* 49.155–180.

BRUCE, GÖSTA. 1977. *Swedish word accents in sentence perspective*. Lund : Gleerup.

BRAUN, BETTINA, and LARA TAGLIAPETRA. 2011. On-line interpretation of intonational meaning in L2. *Language, Cognition and Neuroscience* 26.224–235.

BÜRKI, AUDREY. 2018. Variation in the speech signal as a window into the cognitive architecture of language production. *Psychonomic Bulletin and Review* 25.1973–2004.

CALHOUN, SASHA. 2010. The centrality of metrical structure in signaling information structure: a probabilistic perspective. *Language* 86.1–42.

CARBARY, KATHLEEN; MEREDITH BROWN; CHRISTINE GUNLOGSON; JOYCE M. MCDONOUGH; ALESKANDRA FAZLIPOUR; and MICHAEL K. TANENHAUS. 2015. Anticipatory deaccenting in language comprehension. *Language, Cognition and Neuroscience* 30.197–211.

CHEN, AOJU, and LOU BOVES. 2018. What's in a word: Sounding sarcastic in British English. *Journal of the International Phonetic Association* 48.57–76.

CRUTTENDEN, ALAN. 1997. *Intonation* (2nd ed.). Cambridge: Cambridge University Press.

CRYSTAL, DAVID. 1969. Review of *Intonation and grammar in British English*, by M. A. K. Halliday. *Language* 45.378–393.

DILLEY, LAURA C., and MARA BREEN. 2022. Commentary on Chapter 4: An Enhanced Autosegmental-Metrical Theory (AM+) facilitates phonetically transparent prosodic annotation. In Barnes & Shattuck-Hufnagel, 182–203.

FLETCHER, JANET. 2010. The prosody of speech: timing and rhythm. In Hardcastle et al., 523–602.

FLETCHER, JANET. 2015. Intonation and prosody in Dalabon. In Jun, 252–272.

FLETCHER, JANET; HYWEL STOAKES; RUTH SINGER; and DEBORAH LOAKES. 2016. Intonational correlates of subject and object realization in Mawng (Australian). *Proceedings of the 8th International Conference on Speech Prosody (Speech Prosody 2016)*, Boston, 188–192. Online: https://www.isca-speech.org/archive/pdfs/speechprosody_2016/fletcher16_speechprosody.pdf.

FOX, ROBERT ALLEN, and EWA JACEWICZ. 2009. Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America* 126.2603–2618.

FREEMAN, VALERIE. 2021. Vague eggs and tags: Prevelar merger in Seattle. *Language variation and change* 33(1).57-80.

FROTA, SÓNIA. 2014. The intonational phonology of European Portuguese. In Jun, 6–42.

FUJISAKI, HIROYA. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. *The Production of Speech*, ed. by Peter MacNeilage, 39–55. Heidelberg: Springer.

FUNG, ROXANA S.Y., and CHRIS K. C. LEE. 2019. Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception. *Journal of the Acoustical Society of America* 146 (5): EL424–EL430. https://doi.org/10.1121/1.5133661.

GEORGAKOPOULOS, THANASIS, and STAVROS SKOPETEAS. 2010. Projective vs. interpretational properties of nuclear accents and the phonology of contrastive focus in Greek. *The Linguistic Review* 27.319–346.

GRABE, ESTHER; GREG KOCHANSKI; and JOHN COLEMAN. 2007. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech* 50.281–310.

GRICE, MARTINE; SIMON RITTER; HENRIK NIEMANN; and TIMO B. ROETTGER. 2017. Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics* 64.90–107.

GRICE, MARTINE; ALEXANDRA VELLA; and ANNA BRUGGEMAN. 2019. Stress, pitch accent, and beyond: Intonation in Maltese questions. *Journal of Phonetics* 76.100913. Online: https://doi.org/10.1016/j.wocn.2019.100913.

GRICE, MARTINE; SIMON WEHRLE; MARTINA KRÜGER; MALIN SPANIOL; FRANCESCO CANGEMI; and KAI VOGELEY. 2023. Linguistic prosody in autism spectrum disorder—An overview. *Language and Linguistics Compass* 17(5): e12498. Online: https://doi.org/10.1111/lnc3.12498

GRYLLIA, STELLA; MARY BALTAZANI; and AMALIA ARVANITI. 2018. The role of pragmatics and politeness in explaining prosodic variability. *Proceedings of the 9th International Conference on Speech Prosody (Speech Prosody 2018),* Poznań, 158–162. Online: https://www.isca-speech.org/archive/pdfs/speechprosody_2018/gryllia18_speechprosody.pdf.

GU, WENTAO; KEIKICHI HIROSE; and HIROYA FUJISAKI. 2007. Analysis of tones in Cantonese Speech based on the command-response model. *Phonetica* 64.29–62.

GUBIAN, MICHELE; FRANCESCO TORREIRA; and LOU BOVES. 2015. Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics* 49.16–40.

GUBIAN, MICHELE; JONATHAN HARRINGTON; MARY STEVENS; FLORIAN SCHIEL; and PAUL WARREN. 2019. Tracking the New Zealand English NEAR/SQUARE merger using functional principal components analysis. *Proceedings of Interspeech 2019*. 297–300. Online: https://www.isca-speech.org/archive/pdfs/interspeech_2019/gubian19_interspeech.pdf

GUSSENHOVEN, CARLOS. 1992. Illustrations of the IPA: Dutch. *Journal of the International Phonetic Association* 22 (1-2). 45–47.

GUSSENHOVEN, CARLOS. 2004. *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

GUSSENHOVEN, CARLOS, and MARCO VAN DE VEN. 2020. Categorical perception of lexical tone contrasts and gradient perception of the statement-question intonation contrast in Zhumadian Mandarin. *Language and Cognition* 12.614–648.

HALLIDAY, M. A. K. 1967. *Intonation and grammar in British English*, The Hague: Mouton.

HARDCASTLE, WILLIAM J.; JOHN LAVER, and FIONA E. GIBBON (eds.) 2010. *The Handbook of Phonetic Sciences*. Hoboken, NJ: Wiley-Blackwell.

HAYES, BRUCE. 1994. 'Gesture' in prosody: Comments on the paper by Ladd, *Papers in Laboratory Phonology* III, ed. by Patricia Keating, 64–75. Cambridge: Cambridge University Press.

HELLBERND, NELE, and DANIELA SAMMLER. 2018. Neural bases of social communicative intentions in speech. *Social Cognitive and Affective Neuroscience 13*.604–615.

HIRSCHBERG, JULIA. 2004. Pragmatics and intonation. *The handbook of pragmatics*, ed. by Larry R. Horn and Gregory Ward, 515–537. London: Blackwell.

HIRST, DANIEL, and ALBERT F. DI CRISTO. 1998. A survey of intonation systems. *Intonation systems: A survey of twenty languages*, ed. by Daniel Hirst & Albert F. Di Cristo, 1–44. Cambridge University Press.

HENRIKSEN, NICHOLAS. 2013. Style, prosodic variation, and the social meaning of intonation. *Journal of the International Phonetic Association* 43.153–193.

HOLLIDAY, NICOLE R. 2019. Variation in question intonation in the corpus of regional African American language. *American Speech* 94.110–130.

HU, NA, and AMALIA ARVANITI. 2023. Individual variation in the use of non-F0 cues in the realization of accentual contrasts. Paper presented at the Acoustical Society of America Fall meeting, Sydney.

HUALDE, JOSÉ IGNACIO, and PILAR PRIETO. 2016. Towards an International Prosodic Alphabet (IPrA). *Laboratory Phonology* 7(1).5. Online: http://doi.org/10.5334/labphon.11.

INTERNATIONAL PHONETIC ASSOCIATION. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, U.K: Cambridge University Press.

JOHNSON, PAUL C. D. 2014. Extension of Nakagawa; and Schielzeth's R^2_{GLMM} to random slopes

JUN, SUN-AH (ed) 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

JUN, SUN-AH (ed) 2014. *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

JUN, SUN-AH. 2022. ToBI is not designed to be phonetically transparent: Response to Dilley; and Breen's commentary to Jun. In Barnes & Shattuck-Hufnagel, 204–211. Cambridge, MA: The MIT Press.

JUN, SUN-AH, and JIHYEON CHA. 2015. High-toned [il] in Korean: Phonetics, intonational phonology, and sound change. *Journal of Phonetics* 51.93–108.

KARLIN, ROBIN. 2018. *Towards an articulatory model of tone: a cross-linguistic investigation*. PhD dissertation, Cornell University.

KARLIN, ROBIN. 2022. Expanding the gestural model of lexical tone: Evidence from two dialects of Serbian. *Laboratory Phonology* 13(1). Online: https://doi.org/10.16995/labphon.6443.

KATSIKA, ARGYRO. 2016. The role of prominence in determining the scope of boundary-related lengthening in Greek. *Journal of Phonetics* 55.149–181.

KATSIKA, ARGYRO, and KAREN TSAI. 2021. The supralaryngeal articulation of stress and accent in Greek. *Journal of Phonetics* 88.101085.

KATSIKA, ARGYRO; JELENA KRIVOKAPIĆ; CHRISTINE MOOSHAMMER; MARK TIEDE; and LOUIS GOLDSTEIN. 2014. The coordination of boundary tones and its interaction with prominence. *Journal of Phonetics* 44.62–82.

KAWAHARA, HIDEKI; ALAIN DE CHEVEIGNÉ; HIDEKI BANNO; TORU TAKAHASHI; and TOSHIO IRINO. 2005. Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. *Proceedings of Interspeech 2005*. Online: https://www.isca-speech.org/archive/pdfs/interspeech_2005/kawahara05_interspeech.pdf.

KELLEY, MATTHEW C., and BENJAMIN V. TUCKER. 2020. A comparison of four vowel overlap measures. *Journal of the Acoustical Society of America* 147.137–145.

KHAN, ARBAAZ; YI XU; and AYESHA SOHAIL. 2020. Multidimensionality of tone in Pahari. *Lingua* 245: 102923–102923. Online: https://doi.org/10.1016/j.lingua.2020.102923

KIM, JISEUNG; NA HU; RICCARDO ORRICO; STELLA GRYLLIA; and AMALIA ARVANITI. 2023. The phonetics and pragmatics of H* and L+H* in British English. Paper presented at the 2nd International Conference on Tone and Intonation (TAI 2023), Singapore.

KNIGHT, RACHAEL-ANNE. 2008. The shape of nuclear falls and their effect on the perception of pitch and prominence: Peaks vs. plateau. *Language and Speech* 51(3).223–244.

KOTTMANN, TABEA; MAREN WANNER; and KATHLEEN WERMKE. 2023. Fundamental frequency contour (melody) of infant vocalizations across the first year. *Folia Phoniatrica et Logopaedica* 75(3): 177–187.

KRIVOKAPIĆ, JELENA. 2022. Prosody in Articulatory Phonology. In Barnes & Shattuck-Hufnagel, 213–236.

KURUMADA , CHIGUSA, and TIMO B. ROETTGER. 2022. Thinking probabilistically in the study of intonational speech prosody. *Wiley Interdisciplinary Reviews: Cognitive Science* 13(1). e1579. Online: https://doi.org/10.1002/wcs.1579.

KUZNETSOVA, ALEXANDRA; PER B. BROCKHOFF; and RUNE H.B. CHRISTENSEN. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13).1–26.

LADD, D. ROBERT. 2008. *Intonational Phonology*. Cambridge: Cambridge University Press.

LADD, D. ROBERT. 2014. *Simultaneous structure in phonology*. Oxford: Oxford University Press.

LADD, D. ROBERT; INEKE MENNEN; and ASTRID SCHEPMAN. 2000. Phonological conditioning of peak alignment in rising pitch accents in Dutch. *The Journal of the Acoustical Society of America* 107.2685–2696.

LENTH, RUSSEL V. 2021. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.6.0. Online: https://CRAN.R-project.org/package=emmeans

LIBERMAN, MARK Y., and JANET B. PIERREHUMBERT. 1984. Intonational invariance under changes in pitch range and length. *Language sound structure: Studies in phonology presented to Morris Halle*, ed. by Mark Aronoff and Richard T. Oehrle, 157–233. Cambridge, MA: The MIT Press.

LUCERO, JORGE C.; KEVIN G. MUNHALL; VINCENT L. GRACCO; and JAMES O. RAMSAY. 1997. On the registration of time and the patterning of speech movements. *Journal of Speech, Language, and Hearing Research* 40.1111–1117.

MACHAČ, PAVEL, and RADEK SKARNITZL. 2009. *Principles of Phonetic Segmentation*. Praha: Epocha Publishing House.

McMurray, Bob, and Allard Jongman. 2011. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review* 118.219–246.

Martin, Andrea E. 2016. Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology* 7.120. Online:10.3389/fpsyg.2016.00120.

Maskikit-Essed, Raechel, and Carlos Gussenhoven. 2016. No stress, no pitch accent, no prosodic focus: The case of Ambonese Malay. *Phonology* 33.353–389.

Nadeu Marianna, and Pilar Prieto. 2011. Pitch range, gestural information and perceived politeness in Catalan. *Journal of Pragmatics* 43(3).841–54.

Nakagawa, Shinichi, and Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2). 133–142.

Nakai, Satsuki, and James M. Scobbie. 2016. The VOT category boundary in word-initial stops: Counter-evidence against rate normalization in English spontaneous speech. *Laboratory Phonology* 7(1).13. Online: http://doi.org/10.5334/labphon.49.

Nespor, Marina, and Angela Ralli. 1996. Morphology-phonology interface: Stress domains in Greek compounds. *The Linguistic Review* 16.357–382.

Niebuhr, Oliver. 2012. At the edge of intonation: The interplay of utterance-final F0 movements and voiceless fricative sounds in German. *Phonetica* 69.7–21.

Niebuhr, Oliver; Mariapaola D'Imperio; Barbara Gili Fivela; and Francesco Cangemi. 2011. Are there “Shapers” and “Aligners”? Individual differences in signalling pitch accent category. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*, Hong-Kong, 120–123. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/SpecialSession/Session4/Niebuhr/Niebuhr.pdf.

Nolan, Francis. 2022. The rise and fall of the British School of intonation analysis. In Barnes & Shattuck-Hufnagel, 319–349.

O’ Connor, J. D. , and G. F. Arnold. 1973. *Intonation of colloquial English*. London: Longman.

Parker, Ellen M.; Randy L. Diehl; and Keith R. Kluender. 1986. Trading relations in speech and nonspeech. *Perception and Psychophysics* 39.129–142.

Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation*. MIT dissertation.

Pierrehumbert, Janet. 2002. Word-specific phonetics. *Laboratory Phonology VII*, ed. by Carlos Gussenhoven and Natasha Warner, 101–139. Berlin: Mouton de Gruyter.

Pierrehumbert, Janet. 2016. Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics* 2.33–52.

Pierrehumbert, Janet. 2022. Commentary on Chapter 11: Comparing the PENTA Model to Autosegmental-Metrical Phonology. In Barnes & Shattuck-Hufnagel, 408–424.

Pierrehumbert, Janet, and Mary Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: The MIT Press.

Pierrehumbert, Janet B. , and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. *Intentions in communication*, ed. by Phillip R. Cohen, Jerry L. Morgan and Martha E. Pollack, 271–311. Cambridge, MA: The MIT Press.

PIKE, KENNETH L. 1945. *The Intonation of American English*. Ann Arbor, MI: U. of Michigan Press.

PRIETO, PILAR. 2014. The intonational phonology of Catalan. In Jun, 43–80.

PRIETO, PILAR, and PAOLO ROSEANO. 2021. The encoding of epistemic operations in two Romance languages: The interplay between intonation and discourse markers. *Journal of Pragmatics* 172.146–163.

PROM-ON, SANTITHAM; YI XU; WENTAO GU; AMALIA ARVANITI; HOSUNG NAM; and DOUGLAS H. WHALEN. 2016. The Common Prosody Platform, (CPP) – where theories of prosody can be directly compared. *Proceedings of the 8th International Conference on Speech Prosody (Speech Prosody 2016)*, Boston. Online: https://www.isca-speech.org/archive/pdfs/speechprosody_2016/promon16_speechprosody.pdf.

R CORE TEAM. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Online: https://www.R-project.org/.

RAMSAY, JAMES. 2005. Functional Data Analysis. *Encyclopedia of Statistics in Behavioral Science* edited by Brian Everitt and David Howell. Online: https://doi.org/10.1002/0470013192.bsa239.

RAMSAY, JAMES O., and BERNARD W. SILVERMAN. 2005. *Functional Data Analysis*. New York, NY: Springer. Online: https://doi.org/10.1007/0-387-22751-2_8

RAMSAY, JAMES O.; SPENCER GRAVES; and GILES HOOKER. 2020. fda: Functional Data Analysis. R package version 5.1.9. Online: https://CRAN.R-project.org/package=fda

RATHCKE, TAMARA. 2017. How truncating are 'truncating languages'? Evidence from Russian and German. *Phonetica* 73.194–228.

RAVIV, LIMOR; GARY LUPYAN; and SHAWN C. GREEN. 2022. How variability shapes learning and generalization. Trends in Cognitive Sciences 26.462–483.

REPP, BRUNO H. 1982. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin* 92.81–110.

ROESSIG, SIMON, and DORIS MÜCKE. 2019. Modeling Dimensions of Prosodic Prominence. *Frontiers in Communication* 4. Online: https://doi.org/10.3389/fcomm.2019.00044.

ROETTGER, TIMO B. 2017. *Tonal placement in Tashlhiyt: How an intonation system accommodates to adverse phonological environments*. Berlin: Language Science Press. Online: 10.5281/zenodo.814472.

ROETTGER, TIMO B. , and MARTINE GRICE. 2019. The tune drives the text - Competing information channels of speech shape phonological systems. *Language Dynamics and Change* 9(2).265–298.

SCHERTZ, JESSAMYN, and EMILY J. CLARE. 2020. Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive science* 11(2).e1521. Online: https://doi.org/10.1002/wcs.1521.

SCHMIDT, ELAINE; ANA PÉREZ; LUCA CILIBRASI; and IANTHI TSIMPLI. 2020. Prosody facilitates memory recall in L1 but not in L2 in highly proficient listeners. *Studies in Second Language Acquisition* 42(1). 223–238.

SCOBBIE, JAMES M. 2007. Interface and overlap in phonetics and phonology. *The Oxford Handbook of Linguistic Interfaces*, ed. by Gillian Ramchand; and Charles Reiss. Oxford: Oxford University Press. 17-52.

SHUE, YEN-LIANG. 2010. *The voice source in speech production: Data, analysis and models*. UCLA dissertation.

SHUE, YEN-LIANG; PARTICIA KEATING; CHAD VICENIK; and KRISTINE YU. 2011. VoiceSauce: A program for voice analysis. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*, Hong-Kong, 1846–1849. Online: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Shue/Shue.pdf.

SILVERMAN, KIM, and JANET PIERREHUMBERT. 1990. The timing of prenuclear high accents in English. *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, ed. by John Kingston and Mary Beckman, 72–106. Cambridge: Cambridge University Press.

SMILJANIC, RAJKA. 2006. Early vs. late focus: pitch-peak alignment in two dialects of Serbian and Croatian. *Laboratory Phonology 8*, ed. by Louis Goldstein, Douglas H. Whalen, and Catherine T. Best, 495–518. Berlin: Mouton de Gruyter.

STANLEY, JOSEPH A., and BETSY SNELLER. 2023. Sample size matters in calculating Pillai scores. *Journal of the Acoustical Society of America* 153 (1).54–67.

STEEDMAN MARK. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry* 31.649–689.

STEFFMAN, JEREMY, and SUN-AH JUN. 2019. Perceptual integration of pitch and duration: Prosodic and Psychoacoustic influences in speech perception. *Journal of the Acoustical Society of America* 146(3).EL251– EL257.

STEVENS, KENNETH N. , and SAMUEL JAY KEYSER. 1989. Primary features and their enhancement in consonants. *Language* 65.81–106.

STRANGE, WINIFRED; ANDREA WEBER; ERIKA S. LEVY; VALERIY SHAFIRO; MIWAKO HISAGI; and KANAE NISHI. 2007. Acoustic variability within and across German, French, and American English vowels: Phonetic context effects. *The Journal of the Acoustical Society of America* 122**.**1111-1129.

'T HART, JOHAN; RENÉ COLLIER; and ANTONIE COHEN. 1990. *A perceptual study of intonation: An experimental phonetic approach to speech melody*. Cambridge: Cambridge University Press.

YU, ALAN C. L. 2023. Tonal context influences tone-duration interaction: Evidence from Cantonese. *JASA Express Letters* 3.035203. Online: https://doi.org/10.1121/10.0017517

WANG, LEI; CARLOS GUSSENHOVEN; and JIE LIANG. 2020. How to pronounce a low tone: A lesson from Kaifeng Mandarin. *Journal of the International Phonetic Association* 50.199–219.

WU, YARU; MARTINE ADDA-DECKER; and LORI LAMEL. 2023. Mandarin lexical tone duration: Impact of speech style, word length, syllable position and prosodic position. *Speech Communication* 146.45–52.

ZHANG, YUBIN, and JAMES KIRBY. 2020. The role of F0 and phonation cues in Cantonese low tone perception. *The Journal of the Acoustical Society of America* 148, 1.EL40–EL45. Online: https://doi.org/10.1121/10.0001523.

ZHENG, XIAOJU. 2006. Voice quality variation with tone and focus in Mandarin. *Proceedings of TAL 2006–International Symposium on Tonal Aspects of Languages. La Rochelle,* 132–136. Online: https://www.isca-speech.org/archive/pdfs/tal_2006/zheng06_tal.pdf