Embarking on PoLaR Explorations: A Framework for Intonational Annotation and Analysis

Byron Ahn,
Nanette Veilleux,
Stefanie Shattuck-Hufnagel,
and Alejna Brugos

draft
November 2022

Contents

List of Figures

List of Tables

Chapter 1

Background, Motivation, and Overview

1.1 Introduction to Prosody and Prosodic Annotation

In addition to being formed of words, spoken utterances contain a wide range of other information about timing, intonation, prominence, phrasing, voice quality, rhythm, etc., often collectively called spoken prosody. (See Ladd 2008, Beckman & Venditti 2011, and Barnes & Shattuck-Hufnagel *forthcoming* for some broad overviews.) These aspects of an utterance are sometimes called supra-segmental, because they can span regions larger than a single phonemic segment (i.e., a single consonant or vowel). (See Lehiste 1970 for extensive discussion.)

In a language like English, two major categories of prosodic structure concern **prominence** (related to notions of accent, stress, focus, emphasis, etc.) and **phrasing** (related to notions of grouping, disjuncture, pauses, etc.). In turn, both prominence and phrasing correlate with changes in **pitch** (related to notions of f0, tone, intonation, etc.). Speakers of English modulate these and other prosodic aspects of speech and thereby signal distinctive pragmatic, semantic, syntactic, or morphological information. In order to study these phenomena, linguists and speech scientists of many types are interested in annotating the prosodic structure of utterances.

As an example of the effect of prosodic manipulation on linguistic structures and meanings that speech scientists and linguists have been interested in, consider the English string "*Steve or Sam and Bob will come*". As discussed in Lehiste 1973 (also Price et al. 1991, Veilleux et al. 2006), manipulating the supra-segmentals that signal prominence and grouping in this sentence can change its fundamental meaning. In the following two pronunciations, capitalization indicates prominence and commas indicate phrasing.

1. STEVE, or Sam and BOB, will come.
2. Steve or SAM, and BOB, will come.

This simple manipulation of prominence and phrasing highlights the linguistic importance of prosody. Each of these two realizations of the same string (which are two of many possibilities) yields a fundamentally different structure and interpretation: the former is unclear about whether one or two people will come (Steve alone, or Sam and Bob together), while the latter more clearly communicates that two people will come and one of them will be Bob. Under-

standing this kind of prosodic patterning can be useful in a wide variety of domains, e.g., in formulating the linguistic grammar, modelling human speech production and perception, mapping prominence and grouping patterns to meaning differences, understanding the effects of prominence and grouping on the pronunciation of words, developing better-performing algorithms for automatic speech synthesis, recognition and translation, and improving understanding of speech disorders that involve prosody. To address these goals, researchers in intonation (and prosody more generally) need to be able to systematically annotate a variety of prosodic differences, in ways that go beyond laboratory examples and stylized productions, and capture aspects of the phonetic implementation of phonological prosodic contrasts.

1.1.1 Pitch Cues to Prominence and Phrasing

Though prominence and phrasing are abstract concepts, manipulation of the intonational acoustics of an utterance can provide strong cues as to which elements are prominent and where phrase boundaries exist.[1] A particularly strong set of cues comes from changes in perceived pitch that are caused by changes in the frequency of vibration of the vocal folds (this vibration rate is often called "f0", for "fundamental frequency").[2] In terms of the meaning of a sentence, intonational differences can play a key role (as exemplified by sentences 1 and 2 above).

However, the relationship between pitch and meaning can be complex. For example, high pitch (acoustically measured as f0) can signal that a particular word is meaningfully prominent in English; however, 'high pitch' can map onto a wide range of f0 values in the acoustics, depending on context. This is because what counts as 'high' in one context, might be much higher/lower (in terms of f0 values) than what counts as 'high' in another. Moreover, it's not just high f0 values that signal that a word is prominent in English; prominence can also be signaled by an f0 pattern that is low, rising, falling, or etc. In other words, there is no fixed 1-to-1 relationship between an f0 value and prominence.

In addition to signalling prominence, a high f0 can also be used to mark a phrase boundary, as in the pitch rise often heard on the final syllable of certain kinds of questions in English, such as "*Is it raining yet?*" Here, when f0 rises to a high value at the end of *yet* it does not necessarily mark a pitch-accented word; in fact, in perhaps most pronunciations of this question, *yet* is not a phrasally-prominent word. Instead, a high f0 on *yet* can signal the presence of a phrase boundary following it.

The paragraphs above reveal that high f0 values can serve as cues to both prominence and phrase boundaries. Moreover, it is not always straightforward to determine whether a high f0 region serves as a cue to prominence or phrasing (or neither). Identifying these different **types** of f0 patterns (prominence- vs. phrasing-related) requires a theory (an intonational phonology), training, and often extensive practice. At the same time, one could –without knowing whether f0 changes are prominence- or phrasing-related– annotate where significant changes in f0 trajectory (realized as, e.g., peaks or valleys) occur. PoLaR is designed with this goal in mind: It allows labellers to annotate perceptually-significant f0 changes separately from prosodic events like phrase boundaries and prominence, while still permitting annotation of these relationships where they are perceived to exist.

[1] The concept of prominence has been defined in a variety of ways, as required by different disciplines. For further discussion see Gussenhoven 2015, Wagner et al. 2015.

[2] Note that the cues to phrasing and prominence are *by no means* restricted to the acoustics of f0. Speakers also manipulate dimensions such as duration, amplitude and voice quality (phonation quality) to signal prosodic structure. For some further discussion, see section **?REF?**.

To summarize, prosody encompasses many different aspects of the speech signal – beyond words and their phonological representations as sequences of consonants and vowels. Here we have focused on the intonational aspects of prosody, noting that PoLaR allows novice and advanced labellers to contribute differently to its annotation, according to their level of knowledge and their goals. This feature distinguishes PoLaR from some other prosodic annotation systems, which may have more fixed requirements based on a particular phonological theory of prosody.

1.1.2 Terminology: f0, pitch, intonation, and prosody

Before continuing, we will clarify our working definitions for some terms that are used throughout this monograph. We start with **f0** and **pitch**, because these two ideas are often conflated, especially in more casual discussion, even though there is an important difference between them. Fundamental frequency (f0) in speech is directly related to the rate of vocal fold vibration, and is estimated by signal detection algorithms in software like Praat. That is, what Praat calls the "pitch track" (shown in blue in the figures throughout this monograph) is more precisely an (estimated) f0 contour. On the other hand, pitch is not directly measurable - it is a psycho-perceptual phenomenon. As such, pitch only exists in the mind of a listener. To describe this another way, if there were a speaking event such that no one heard the speech, the utterance would have an f0 contour but no pitch, because pitch does not exist outside of the minds of a listener.

Both f0 and pitch are dynamic, changing in patterned ways over the course of an utterance. These dynamic changes are often visualized as a graph, where the x-axis represents time and the y-axis represents f0 values; the visualizations of f0 changes are correspondingly called **f0 contours** (a.k.a. "f0 tracks"). On the other hand, a more abstract representation of how a listener perceives pitch changes over time (e.g., a visualization like a straight line approximation) is called a **pitch contour**. Thus, an f0 contour is a description of changes in the *f0 values* over time, while a *pitch* contour is a description of changes in the perceived pitch over time. (Note that since, in our view, pitch does not exist without a listener with a mind to represent it, a pitch contour also does not exist without a listener with a mind.) Abstracting further over these contours, using discrete grammatical objects, produces what we call the **intonational contour**, which is an abstract sequence of pitch events (targets) that can occur over time in a spoken utterance.

This brings us to the term **intonation**, which we take to refer to the portion of phonetics/phonology that deals in describing/modelling patterns in pitch in linguistic utterances. To do so, intonation must make reference to various other aspects of phonetics and phonology, including other aspects of **prosody**. We take prosody to refer to the portion of phonetics/phonology that deals in describing/modelling patterns in suprasegmentals (i.e., patterns in the signal that can extend across multiple segments; see Lehiste 1970 for more discussion) in linguistic utterances. In other words, these definitions treat intonation as a subset of prosody. (At this point, it is worth mentioning that there are blurred lines in any conceptual distinctions here. The distinctions are blurry in part because the concepts are not discrete, because they interact with one another, and because colloquial usages of the terms are not always consistent.)

A summary of these working definitions is provided in the table below.

f0	a physical measure directly related to rate of vibration of the vocal folds, as reflected in the acoustic signal or articulatory measures
pitch	an abstract psycho-perceptual phenomenon related to f0 (*requires a listener with a mind*)
f0 contour	a description of changes in the f0 values in an utterance over time
pitch contour	a description of changes in pitch over time (*represents events in a listener's mind*)
intonational contour	an abstract sequence of pitch events over time (*requires a grammar*)
intonation	the arm of phonetics/phonology dealing with pitch patterns
prosody	the arm of phonetics/phonology dealing more broadly with suprasegmental patterns

Table 1.1: Our working definitions for some commonly used terminology.

1.2 Motivation for PoLaR

Those who are new to intonation and prosody should ***feel free to skip this section***. It is mostly aimed at positioning PoLaR in the literature on prosody and prosodic annotation. It has been written for an audience that is at least somewhat familiar with the issues of intonation (and prosody and suprasegmentals, more generally) as well as issues of already-established systems of annotating intonation.

1.2.1 PoLaR Influences: Some Approaches to Prosodic Labelling

Systems for labelling prosodic information can vary from one to another, even in ways as fundamental as which aspects of the signal are attended to or the number of different symbols in the annotation 'alphabet'. This holds even for annotation within a single language like English, and even for a single idealized variety of English, such as mainstream US English. In developing the PoLaR system, we have made extensive use of some of the concepts and ideas that have also been components of other labelling traditions:

- American structuralism (e.g., Pike 1945, Trager & Smith 1951),
- the British school (e.g., Crystal 1969, O'Connor & Arnold 1973),
- the Dutch IPO model (e.g., 't Hart et al. 1990), and
- the Autosegmental-Metrical framework (e.g., Pierrehumbert 1980, Beckman & Ayers 1997, Grabe et al. 2001, Hualde & Prieto 2016, Dilley & Breen 2018),
- among others (e.g., Hirst 2007, Taylor 1998, Xu 2012).

(For further description of past prosodic models and annotation systems, please see, e.g., Roach 1994, Ladd 2008 Chapters 1 and 2, Féry 2017 Chapter 5, Barnes & Shattuck-Hufnagel *forthcoming*.) That said, **no** familiarity with these systems is required in order to learn and apply the basic aspects of PoLaR annotation.

1.2.2 Context and Motivating Questions

PoLaR was developed in the context of many discussions over long periods of time, in which labellers well-versed in intonational annotation grappled with how to decide on the appropriate intonational label for certain contours (particularly in English), and in particular cases where the crucial differences appeared to involve considerations that are not always explicitly acknowledged. In particular, the present authors have been involved in the development, instruction, and maintenance of the MAE_ToBI system (*M*ainstream *A*merican English Tones and Break Indices; Beckman & Hirschberg 1994, Beckman & Ayers 1997, Beckman et al. 2005 currently embodied in MIT's Open Courseware system [link]). While committed to the development of PoLaR, the authors remain interested and invested in ToBI annotation systems for labelling phonological categories; we believe the systems are complementary, and not in competition. ToBI is a phonological annotation system, for transcribing intonational categories. It was developed within the framework of AM (Autosegmental-Metrical) phonology (as in Pierrehumbert 1980, Ladd 2008, Arvaniti & Fletcher 2020), which distinguishes different levels of prosodic phrases (e.g., "Intermediate Phrases" and "Intonation Phrases"), as well as different types of pitch movements (those associated with stressed syllables [e.g., "Pitch Accents"] and those associated with prosodic phrase edges [e.g., "Phrase Accents and Boundary Tones"]). While PoLaR can distinguish such objects, it doesn't require that its labellers commit to any particular phonological analyses. In this sense it contrasts with ToBI, in which all phonological categories of pitch are annotated as either categorically high (H) or low (L), following Pierrehumbert 1980).

These discussions reflected the sense that, while existing (AM-based) phonological models of English intonation (e.g., MAE_ToBI) are well-suited to capture many phonological aspects of the intonation system, they purposefully avoid capturing the finer details of intonation contours (and other aspects of prosody). Because these details may be systematically determined, and furthermore may possibly signal additional categories and meanings, it became clear that a way needed to be found to permit their annotation. In particular, three questions emerged from these extensive discussions that have ultimately shaped the PoLaR system:

Three Motivating Questions

Question 1: Which phonetic cues does/should a labeller attend to in labelling phonological categories?

Question 2: What is the range of possible suprasegmental phonetic implementations for a given phonological category?

Question 3: What are the ways in which prosody signals meaning, inclusive of and perhaps even beyond the phonological categories of current systems?

Question 1) **Which cues?** Labellers using phonological systems must still attend to acoustic cues, in order to determine which phonological label to use. At the same time, different labellers may make use of different cues and weight them differently (or even disregard them completely), leading to different phonological labels for the same observed set of cues. One of the motivations for developing PoLaR was to facilitate discussions of how each labeller interprets cues, by having them explicitly annotate the cues they attend to – in PoLaR's case, the intonational cues. (See **?REF?** in Chapter **?REF?** for further discussion.)

Question 2) What range of surface forms? There is still much to be learned about the range of surface forms that can be used to signal a particular phonological category of pitch accent or edge tone - even for well-studied languages like English. PoLaR adds explicit focus on the acoustic details of the signal, so that a corpus with both PoLaR labels and more complex phonological (e.g., ToBI, RaP, IViE) labels will provide an inventory of surface phonetic realizations of each proposed phonological category.

Question 3) Which meanings? Despite decades of study of how prosody contrastively conveys meaning, it is not entirely certain that any existing phonological system of prosodic annotation captures all of the phonological categories of the prosodic system. For example, developments in the literature suggest that certain aspects of English intonational contours currently not captured by MAE_ToBI labels may be particularly relevant for signaling semantic-pragmatic meanings (e.g., range size [cf. Ladd 1994] and or certain boundary-related movements [cf. Ahn et al. 2016]), beyond those signaled by the presence and type of pitch accents and hierarchical phrase boundaries. It is important to understand the ways in which meaning is affected, so as to better understand which acoustic changes are categorical, in a phonological sense.

To address these motivating questions, PoLaR provides tools for the annotation of an utterance's acoustic qualities (targeting its prosodic phonetics) as well as some fundamental abstract aspects of its prosodic categories (targeting its prosodic phonology). PoLaR has been designed so that the task of labelling is not burdensome to the labeller (in a way that is especially useful to the novice). One way that this has been achieved is by designing the labels so that acoustic cues and abstract categories can be labelled separately from one another. Another way that this has been achieved is that the categories invoked are rather abstract (e.g., "prominent") are kept to a minimal number, allowing a degree of neutrality with respect to specifics of the prosodic phonology of the language. At the same time, PoLaR is also useful for those with experience in intonational analysis and theory: the PoLaR Advanced labels permit the annotation of which phonetic details are (in the judgment of the labeller) related to the phonological categories of phrase-level prominences (pitch accents) and boundaries (edge tones).

The guidelines chapters of this monograph focus on the annotation of intonational phonetic details in particular (via the Points, Levels, and Ranges tiers), and so note that whenever we say "phonetics" or "acoustics" here, we primarily are referring to intonational phonetics and acoustics. However, the annotation framework we use with PoLaR gives us the ability to expand annotation methods to similarly capture other domains of phonetic cues (timing, amplitude, phonation, etc.) that are relevant to prosodic structure. (We return to how to extend PoLaR in Chapter **?REF?** .)

1.3 The PoLaR system

1.3.1 PoLaR Tiers and A Labelled Example

Some primary goals of PoLaR are:

1. to annotate a wider array of prosodically relevant features of speech than is possible in existing systems;
2. to isolate different prosodically relevant aspects of the speech signal from one another; and

3. to make the labelling task easier, by requiring fewer phonological decisions.

An example annotated according to the PoLaR labelling guidelines is given in Figure 1.1.

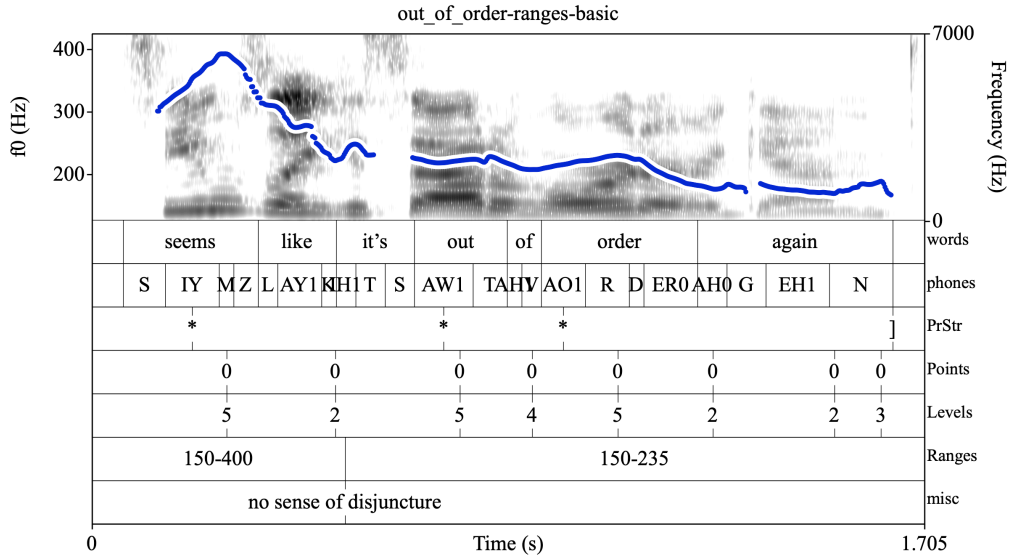

Figure 1.1: A recording annotated with Basic PoLaR labels. Note that the Phones tier is created automatically by the Montreal Forced Aligner (McAuliffe et al. 2019).[3]

There are four core tiers of prosodic annotation (the third through sixth tiers in Figure 1.1 above), described in these guidelines: three acoustic tiers (Pitch **Po**ints, Scaled **L**evels, **a**nd **R**ange Domains), and one phonological tier (Prosodic Structure).

1. The Prosodic Structure (PrStr) tier is a point tier, and each label indicates the presence of a perceived prosodic prominence or a perceived prosodic phrase boundary. Prominence labels are placed in the middle of the vocalic nucleus of a prominent syllable, and boundary labels at the end of the last word of a prosodic phrase.
2. The Pitch **Po**ints tier is a point tier, and each label corresponds to a turning point that the labeller observes in the f0. "Turning point" refers to any point in the f0 curve that looks to be a place where the f0 curve's slope changes significantly (i.e., peaks, valleys, and the edges of plateaus).[4] PoLaR also permits labellers to *optionally* annotate, for each f0 turning point, which (type of) phonological object from the Prosodic Structure tier it is associated with. This tier is the aspect of PoLaR that requires the most substantial discussion, in part to establish which slope changes are significant, what types of 'decoy' or apparent f0 turning points can be ignored, and how missing turning points can be inferred; see section **?REF?** for this discussion.
3. The Scaled **L**evels tier is a point tier, and has a 1-to-1 relationship with the Points tier, in terms of the number and time alignment of annotations. That is, for each point in the

[3]The inclusion of a Phones tier does not reflect the authors' commitment to the idea that phones have representational reality as bounded linguistic constituents.

[4]PoLaR emphasizes the labelling of f0 turning points, because they are important aspects of an f0 contour, but this does not imply a commitment to an equivalence between turning points and intonational targets. See section **?REF?** on Tonal Center of Gravity for further discussion.

Points tier, a point is added to the Levels tier, and that Levels tier object is labelled with a numerical value that corresponds to where in the current pitch-range (see 4 below) the turning point is. This tier is **automatically derived** from the Pitch Points Tier and the Range Domains tier, using the Levels labeller function of the PoLaR plugin for Praat.

4. The **R**ange Domains tier is an interval tier, which captures a local pitch range for each utterance or section of an utterance. This annotation makes it possible to define the "high" and "low" for a particular stretch of an utterance, which is more explicitly manifested in the labels of the Scaled Levels tier (as in 3 above). For Basic PoLaR labels, the max/min for each Range interval is used to determine the numerical Levels values automatically inserted in the Levels tier.

Some PoLaR labels are phonological in nature (though also somewhat underspecified; e.g., "prominence" or "phrase boundary"), while others are more phonetic (e.g., f0 turning points). Annotating each tier only requires attention to one stream of suprasegmental properties (e.g., the Points tier only identifies f0 turning points); this allows each tier to be annotated on its own.[5]

PoLaR thus *explicitly* annotates both categories (phonology) and acoustic cues (phonetics), but with these streams of information *separated from one another*. We believe that annotating this information separately will reduce confounds in analysis and uncertainty in labellers. As we will see when discussing each tier in more detail, Advanced labels can be used to relate information on the phonological tier to information on the multiple phonetic tiers.

1.3.2 Why These Tiers?

The design choice of all labels and tiers (even these more phonetic ones) is, to some degree, phonologically informed and language-specific. That is, PoLaR labels do not identify just "any old (phonetic) information", but rather information that is likely to be relevant for models of English intonation: e.g., pitch alignment, pitch height, prominence, and pitch range. These tiers and labels were chosen by the designers of PoLaR, based on experience with English intonation, but researchers who want to use PoLaR in another language may need to recalibrate the specific labels and/or tiers that get implemented. Because PoLaR is a framework for exploring the categories and cues to intonational prosody, rather than a fixed set of elements to be labelled, the number and nature of the tiers is extendable to accommodate the needs of particular studies. The following paragraphs review the thinking behind the design choices for each tier.

PrStr:
Following AM theory (cf. Pierrehumbert 1980), we assume that there are different types of intonational events, which are associated with different types of abstract phonological objects. In particular, we assume the now classic view that there are two basic sorts of phonological objects in prosodic structure that have direct influence of intonational contours: those related to intonational prominence and those related to intonational phrasing. (To be clear, the term 'intonational prominence' is meant to invoke a level of 'post-lexical' prominence: prominence higher than the level of lexical stress; cf. Bolinger 1958, Liberman & Prince 1977, and Beckman & Edwards 1994.) By design, all of the labels on this tier avoid indicating anything about

[5] Note that no tier requires bundling information from multiple prosodic domains into a single label (this contrasts with a label like H*, which bundles together prominence, pitch height, f0 turning points, etc.). Some Advanced labels re-connect these separated-out features; this is discussed at length for the Points tier in Chapter **?REF?**.

how they are acoustically realized tonally - even abstractly. For example, differences like H- vs. L-, or H* vs. L* are purposely not captured at all in these labels. (These f0 properties will be captured by other labels on other tiers.) While these phonological objects can be signalled by a variety of cues (including changes in f0, duration, intensity, voice quality, etc.), none of these cues are themselves described by labels on this tier. Instead, what is transcribed is only the labeller's ***perception*** of prominence and phrasing. In this way, these Prosodic Structure tier labels are intentionally agnostic about the range of potential acoustic realizations of these different phonological objects. This method encodes information similar to that encoded by Cole et al.'s 2014 Rapid Prosody Transcription method (RPT), and was influenced by their proposal. In RPT tasks, listeners mark perceived boundaries and prominences without concern for precisely how they are realized. This means that data gathered with an RPT methodology could be automatically translated into the accent and boundary tone markers on the Prosodic Structure tier. Labelling PrStr is designed to be simpler than other prosodic labelling systems, with the goal of allowing others to more easily understand the original labeller's intentions.

Points:

The f0 turning points in an f0 contour play an important role in many different theories. For example, researchers have attributed a relationship between f0 turning points and phonological elements, either directly (e.g., as peaks, valleys, or anchored elbows; see, e.g., Ladd et al. 1999 and Welby 2006) or indirectly (e.g., as important factors in implementing f0 shape and alignment distinctions, as in the Tonal Center of Gravity work of Barnes et al. 2012 et seqq.). For these reasons, it is useful to determine where they are. Unfortunately, at the moment this cannot be done automatically, but requires human intervention, for several reasons. First, f0 is challenging to track automatically, and there are often "missing" turning points (e.g. during voiceless segments or creaky-voiced regions). Second, it is challenging to determine which turning points are significant, and which should in contrast be regarded as 'decoy' points: either too small to make a perceptible difference, or the result of a tracking error. Thus, human labelling of points defined as significant for perception of intonation is required, and this monograph provides guidance for determining significant turning points, identifying decoys and inferring missing points. See section **?REF?** in Chapter **?REF?** for further discussion.

While it is widely agreed that there is a mapping relationship between the types of objects in our Prosodic Structure tier and the f0 turning points of the Points tier, PoLaR does not commit its labellers to any particular analysis of this relationship. In this way, the labeller need not try to keep the phonological model in mind while labelling the Points tier, nor even be familiar with any phonological model. At the same time, PoLaR provides a way for labellers to annotate the relationship between the two tiers. (How to do this is laid out in Section **?REF?** in Chapter **?REF?**.) In this way, the Points tier can also be used for annotating mappings between acoustic events and phonological objects.

Levels:

The Levels tier allows PoLaR to capture the relative height of a Points tier object (on a scale of 1 to 5). This relative height can be useful for analysis, since a raw f0 value does not by itself indicate whether that value is high or low (in the speaker's current intended range). This is because, as noted earlier, a relatively low f0 in the speaker's full possible f0 range may be functionally/phonologically high if the speaker's current f0 range is low, and vice versa. The Levels tier encodes scaled pitch values for each f0 turning point on the Points tier. That value corresponds to the pitch quintile in which it occurs (1 being the lowest quintile and 5 being the highest), with the boundaries for each pitch quintile being calculated on the basis of the pitch

range annotated in the Ranges tier. Annotators should use the PoLaR plugin for Praat to automatically have Levels annotation added, once Points and Ranges tiers have been annotated. Further discussion can be found in chapter **?REF?**.

Ranges:
The Ranges tier reflects that f0 events are always interpreted within a speaker's range–not only their overall speaking range, but within locally determined ranges. The Ranges tier provides the context in which the levels (i.e., on the Levels tier) reflect individual points on the Points tier. That is, this is used to identify whether an f0 point is "high", "low", or somewhere in between, in the context of a particular utterance or part of an utterance. The Ranges labels require human labellers because we need intuitions on which parts of the pitch are perceived to be H or L in the speaker's range. The Ranges tier can be used to capture and reflect the relations and differences among pitch events, both locally within a range, and across ranges.

Labelling Ranges tiers in this way allows analyses that other AM labelling systems do not: relative heights between pitch ceilings/floors in arbitrarily distant parts of the recording can be compared. In AM labelling systems, the pitch range can only be inferred by looking at the labelling and the recording together, alongside a theoretical model of the relationship between phrasing and acoustic measures (e.g., that new intermediate phrases begin new pitch ranges). Such an inference can lead to problems in cases where the labeller and the reader have different assumptions about the relationship between phrases and pitch ranges. This highlights the PoLaR system's core, laid out in the introduction: it keeps track of information that other intonational annotation systems make use of, but differently from those other systems, it requires that such information be tracked *explicitly*.

1.4 Overview of PoLaR's Advantages

Before delving into the details of the system, we describe here several general points about the advantages of PoLaR. Its primary goal is to identify the melody of a spoken utterance; in this sense it has something in common with the IPO approach (so named for the Dutch the Institute for Perception Research, 'Instituut voor Perceptie Onderzoek'), which produces straight line approximations by connecting turning points ('t Hart et al. 1990), which can serve as a proxy for key aspects of the melody. In particular, PoLaR has been designed to have five useful characteristics: Compatibility, Flexibility, Modularity, Accessibility, Expandability, Crosslinguistic usability, and Explicitness; in addition, it has inspired concomitant development of a useful set of Associated Tools.

Compatibility with other annotation systems / prosodic analyses: PoLaR works well with other labelling tools and systems which have different goals, and its use alongside other annotation systems is encouraged; PoLaR is not intended as a complete model of spoken prosody. For example, parallel PoLaR and e.g. ToBI[6] labels can be expected to shed light on both the phonemic inventory of a language and the phonetics-phonology interface.

For these reasons, PoLaR is not intended as a replacement for other annotation systems. As such, PoLaR can be seen as a supplement to existing systems (such as the ones mentioned in

[6]ToBI annotation systems exist for a number of languages and varieties; see Jun 2005, 2014 for works describing ToBI systems for a number of languages.

Section 1.2.1). At the same time, it can stand alone, and PoLaR labellers need not have any familiarity with other prosodic annotation systems.

Flexible for different research goals: The PoLaR system, which builds on existing frameworks and labelling systems, was developed to enable both (1) more detailed descriptions of languages with well-studied intonational phonology, in particular with reference to the capture of acoustic details of intonation for which the linguistic relevance has not yet been determined, and (2) the annotation of phonetic patterns in languages, dialects, or varieties whose phonology has not yet been explored, as a step toward understanding the intonational grammar. Its minimal invocation of language-specific phonology adopts prominence and boundary locations from AM theory, and it focuses on acoustic characteristics that are, according to human judgment, relevant for linguistic signalling (Barnes & Shattuck-Hufnagel Forthcoming).

At its core, a PoLaR annotation is a phonologically-informed (but maximally theory-neutral) labelling of intonational acoustic-phonetic cues. This description brings to the forefront the fact that PoLaR labels are neither purely phonetic, nor purely phonological; instead, they are intended to capture phonologically relevant acoustic aspects of the speech signal. Thus, what is perhaps most important here is the separation between labels for phonological objects from phonetic labels of the acoustic characteristics that serve as cues to those objects, as well as the separation of different acoustic cues each to its own tier, and explicit labelling of more of these acoustic characteristics. We believe the labels that we provide below for each of the proposed labelling tiers are a good starting point for US English varieties, but exemplify what PoLaR annotation can do for any language or variety.

Modularity / "Unbundled" labels: In PoLaR, acoustic-phonetic cues and phonology are annotated separately. (And the phonological labelling is minimal, specifying (in its basic form) only the location of prominences and boundaries.) This reflects design principle: PoLaR **disentangles different types of information** as much as possible - isolating different components of prosody on different tiers of annotation. This unbundling facilitates decision-making during labelling, by requiring only minimal phonological awareness on the part of the labeller. (This stands in contrast with phonological labels that bundle together prominence, pitch alignment and scaling, etc.) This unbundling allows PoLaR to be annotated one at a time (at least with the Basic labels in Ch. **?REF?**), without the need to consider the labels on other tiers - this allows for a 'divide and conquer' approach to the labelling task, in which individuals can specialize in specific tasks, in an assembly line model.

At the same time, for those advanced labellers who are interested in connecting labels to a prosodic theory, PoLaR also provides 'Advanced' labels (Ch. **?REF?**), to allow a labeller to annotate some relationships between tiers. This may facilitate exploration of how prosodic components on one tier relate to components on another. (See Extensibility below.)

Accessibility of use: PoLaR has been designed to be **easy to start using**, with relatively minimal instruction, so that useable data of particular interest to a researcher can be produced quickly. This is in part because PoLaR does not require its labellers to have extensive knowledge of a phonological model of prosody, and in part because the different sets of labels are inherently module (allowing some labellers to be only trained in one area of prosodic labelling).

PoLaR's modularity and lack of reliance on prosodic phonology stands in contrast to existing systems which do not specify how precisely to map labels onto particular sets of cues, and

thus do not enable straightforward investigation of how different speakers (and labellers) use different cues in different contexts, or of what cues speakers use to signal particular contrastive intonational categories.

Because of this accessibility, the tasks of labelling different tiers can be split among different labellers, who can quickly develop expertise in that area. In this way, the first steps of the labelling process are intermediate between the full training process for phonological labelling and the training-free method of Rapid Prosody Transcription (RPT) described in Cole et al. (2014, 2017).

Moreover, while accessible, the system requires that annotators using even the most basic PoLaR labels be explicit about their perception, intuition, and/or analysis - facilitating high-level discussions among more experienced or analysis-oriented users. (In addition, Advanced PoLaR labels allow such analysis-oriented users to systematically transcribe their analyses.)

Expandability of the annotation system: Although this monograph focuses on the *intonational* aspects of spoken prosody, a critical feature of PoLaR's design is that it does not restrict labellers to only annotating this information. To be clear, we mean that PoLaR annotation is broadly intended as a ***framework*** (which is to say it is a way of conceptualizing annotation systems), so that there is not a rigid way for PoLaR annotation to be implemented.

Instead, the particular implementation described in this monograph is intended to be seen as a narrow execution of broader conceptual ideas. A labeller can expand/contract the set of labels on a particular tier or expand/contract the set of tiers that are labelled, so as to adjust what aspects of prosody and cues to prosodic structure are annotated. For example, a labeller may wish to systematically annotate duration, intensity, or phonation cues. Or they may wish to annotate other aspects of prosodic structure, such as lexical stress, footing, etc. This is useful, because tabulating and understanding the individual cues to prosodic prominences and boundaries can provide important insights into how phonetic implementation of a phonological intonation category can vary (Brugos 2015, Brugos et al. 2018).

Labelling projects with different goals can also omit tiers that are judged to be less relevant, allowing a novice labeller to focus on (and become expert in) a particular aspect of the intonation. In this way, more complex labelling tasks can be approached with a divide and conquer strategy, which allows each labeller to become expert and reliable more quickly.

Crosslinguistic usability: PoLaR is designed to be usable for any language as well as for any speech style, register, or dialect *within* a language. PoLaR's focus on acoustic cues and broad phonological categories makes PoLaR useful in the initial stages of exploring a language or dialect whose prosody is under-/un-documented. While some tiers require native speaker intuition (e.g., prominences, boundaries). others could be used without native speaker intuitions (e.g., turning points). Thus, using PoLaR can be a step toward formulating a phonological transcription system of intonationally-undocumented systems. This is critically important, because our current understanding of human speech prosody is based on analysis of a strikingly small proportion of the world's languages, while at the same time, it has become easier to create corpora of recorded utterances in understudied languages for analysis. Thus the time is right for development of a system such as PoLaR which facilitates getting a foothold on the analysis ladder for addressing a new language or dialect.

Beyond work on understudied languages/dialects, PoLaR can also be useful for studying how

suprasegmental cues are used differently across varieties and contexts, within a language/variety. Researchers working with a corpus that is annotated for demographic or contextual information could use PoLaR labels in much the same way as it would be used for crosslinguistic work.[7]

Explicitness of annotation: PoLaR facilitates discussion with respect to differences in labeller intuitions, through its explicit annotation of cues (such as f0 turning points and ranges). Discussions are also facilitated through the relationships described by Advanced labels, which can encode labeller intuitions about the connections across tiers (such as pointers in the Points tier to indicate the direction of affiliation to a phonological boundary or prominence). Similarly, it encourages the development and testing of hypotheses about systematic aspects of these relationships. For example, PoLaR facilitate analyses that reveal not only how each type of prominence and boundary proposed in a phonological theory of intonation can be realized acoustically in different contexts, but also whether the proposed contrastive categories would benefit from extension or revision.

Associated tools: A plugin has been developed that includes a set of scripts to facilitate the labelling process. (As described at several points throughout this monograph; it can be accessed through the OSF repository at `https://doi.org/10.17605/OSF.IO/USBX5`.) This plugin can do things such as helping a labeller determine whether a particular PoLaR label ought to be included, substantially facilitating the resolution of ambiguous cases. Other functions include automatically adding labels on particular tiers on the basis of other labels, and exporting information extracted based on the labels in a format that can be usable for statistical analysis and machine learning. (For more discussion of this plugin, see Chapter 1.2.1.) Finally, the system benefits from integrating with existing tools, such as ones for phone segmentation by forced aligners (e.g. Montreal Forced Aligner, McAuliffe et al. 2019, downloadable from MontrealCorpusTools on Github; or the Penn Forced aligner, Yuan & Liberman 2008, downloadable from the Penn Phonetics Laboratory website).

1.5 Who Will Find PoLaR Useful

This monograph describes a recently developed system for annotating prosody, PoLaR, which is aimed at fulfilling some of the needs described above. PoLaR has been designed with some particular influences and goals in mind, as outlined in the next section. Before getting into the details of those goals, and how they differ from the goals of other annotation systems, we will suggest some circumstances in which we think PoLaR might be particularly useful.

How much experience is required to become a user? PoLaR has been designed to facilitate the labelling process, and to be useful for a wide range of individuals with different levels of experience, backgrounds, and goals. It is designed to require minimal theoretical knowledge, enabling novice labellers to begin producing useful annotations more quickly. It is also designed for research groups where different individuals label different aspects of the signal, in parallel. Lastly, advanced aspects of the system are designed to be useful for advanced labellers or theoreticians, who are interested in the finer details of the prosody.

[7] For those wishing to create their own corpus, see, e.g., Meyerhoff et al. 2011 or Podesva & Sharma 2013 for some discussion of recording this sort of information as well as a description of some best practices.

Which areas of research could benefit? PoLaR was also designed to facilitate making new discoveries, especially with respect to how variation within prosody (and specific prosodic characteristics) relates to variation in a (linguistic and/or extra-linguistic) context. For this reason (alongside its usability for novice labellers), PoLaR will be useful for variationists, sociolinguists, acquisitionists, semanticists/pragmaticists, morphologists, phonologists, etc. –as well as those working at the interfaces– to help understand how prosody relates to these domains. Because it facilitates the annotation of potentially systematic aspects of the acoustic detail of an intonational contour, it is well suited to exploratory analysis of both understudied languages and understudied varieties of familiar languages.

Bibliography

Ahn, Byron & Shattuck-Hufnagel, Stefanie & Veilleux, Nanette. 2016. Evidence and intonational contours: An experimental approach to meaning in intonation. *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology* 189–192.

Arvaniti, Amalia & Fletcher, Janet. 2020. The Autosegmental-Metrical Theory of Intonational Phonology. In Gussenhhoven, Carlos & Chen, Aoju (eds.), *The Oxford Handbook of Language Prosody*, 0. Oxford University Press. doi:10.1093/oxfordhb/9780198832232.013.4

Barnes, Jonathan & Shattuck-Hufnagel, Stefanie. Forthcoming. *Prosodic theory and practice*. Cambridge, MA: MIT Press.

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

Beckman, Mary & Venditti, Jennifer. 2011. Intonation. In *The handbook of phonological theory*, 485–532. John Wiley and Sons. doi:10.1002/9781444343069.ch15

Beckman, Mary E. & Ayers, Gayle M. 1997. Guidelines for ToBI labelling. *The OSU Research Foundation* 3. Available at http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf.

Beckman, Mary E. & Edwards, Jan. 1994. Articulatory evidence for differentiating stress categories. In *Phonological structure and phonetic form: Papers in laboratory phonology iii*, 7–33. Cambridge University Press.

Beckman, Mary E. & Hirschberg, Julia. 1994. The ToBI annotation conventions. ms., Ohio State University.

Beckman, Mary E. & Hirschberg, Julia & Shattuck-Hufnagel, Stefanie. 2005. The original ToBI system and the evolution of the ToBI framework. In Jun, Sun-Ah (ed.), *Prosodic typology: The phonology of intonation and phrasing*, 9–54. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199249633.003.0002

Bolinger, Dwight. 1958. A theory of pitch accent in English. *Word* 14(2-3). 109–149.

Brugos, Alejna. 2015. *The interaction of pitch and timing in the perception of prosodic grouping*: Boston University dissertation. doi:https://hdl.handle.net/2144/14012

Brugos, Alejna & Breen, Mara & Veilleux, Nanette & Barnes, Jonathan & Shattuck-Hufnagel, Stefanie. 2018. Cue-based annotation and analysis of prosodic boundary events. In *Speech Prosody 2018*, 245–249. doi:10.21437/SpeechProsody.2018-50

Cole, Jennifer & Mahrt, Timothy & Hualde, José I. 2014. Listening for sound, listening for meaning: Task effects on prosodic transcription. *Proceedings of Speech Prosody 7* 859–863.

Cole, Jennifer & Mahrt, Timothy & Roy, Joseph. 2017. Crowd-sourcing prosodic annotation. *Computer Speech and Language* 45. 300–325. doi:10.1016/j.csl.2017.02.008

Crystal, David. 1969. *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Dilley, Laura & Breen, Mara. 2018. An enhanced autosegmental-metrical theory (AM^{+}) facilitates phonetically transparent prosodic annotation. *Tonal Aspects of Language 2018* 77–81. doi:10.21437/TAL.2018-14

Féry, Caroline. 2017. *Intonation and prosodic structure*. Cambridge, UK: Cambridge University Press.

Grabe, Esther & Post, Brechtje & Nolan, Francis. 2001. Modelling intonational variation in English: the IViE system. In *Proceedings of Prosody 2000*, 51–58. Poznan, Poland: Adam Mickiewicz University.

Gussenhoven, Carlos. 2015. Does phonological prominence exist? *Lingue e Linguaggio* 14. 7–24.

Hirst, Daniel. 2007. A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the XVIth International Congress of Phonetic Sciences*, 1233–1236. Saarbrucken.

Hualde, José I. & Prieto, Pilar. 2016. Towards an international prosodic alphabet (IPrA). *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1). 5. doi: 10.5334/labphon.11

Jun, Sun-Ah (ed.). 2005. *Prosodic typology*. Oxford: Oxford University Press.

Jun, Sun-Ah (ed.). 2014. *Prosodic typology II*. Oxford: Oxford University Press.

Ladd, D. Robert. 1994. Constraints on the gradient variability of pitch range, or, Pitch level 4 lives! In Keating, Patricia (ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III*, 43–63. Cambridge: Cambridge University Press.

Ladd, D. Robert & Faulkner, Dan & Faulkner, Hanneke & Schepman, Astrid. 1999. Constant ''segmental anchoring'' of f0 movements under changes in speech rate. *The Journal of the Acoustical Society of America* 106. 1543–1554. doi:10.1121/1.427151

Ladd, Robert. 2008. *Intonational phonology*. Cambridge: Cambridge University Press 2nd edn.

Lehiste, Ilse. 1970. *Suprasegmentals*. The MIT Press.

Lehiste, Ilse. 1973. Phonetic disambiguation of syntactic ambiguity. *Glossa* 7(2). 107–122.

Liberman, Mark & Prince, Alan. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8(2). 249–336.

McAuliffe, Michael & Socolof, Michaela & Mihuc, Sarah & Wagner, Michael & Sonderegger, Morgan. 2019. Montreal forced aligner [computer program] (version 1.0.1). Retrieved from `http://montrealcorpustools.github.io/Montreal-Forced-Aligner/`. doi:http://doi.org/10.5281/zenodo.2630943

Meyerhoff, Miriam & Adachi, Chie & Nanbakhsh, Golnaz & Strycharz, Anna. 2011. Sociolinguistic fieldwork. In *The Oxford handbook of linguistic fieldwork*, Oxford University Press. doi:10.1093/oxfordhb/9780199571888.013.0006

O'Connor, J.D. & Arnold, G.F. 1973. *Intonation of colloquial English*. Prentice Hall Press 2nd edn.

Pierrehumbert, Janet. 1980. *The phonology and phonetics of English intonation*: MIT dissertation.

Pike, Kenneth L. 1945. *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.

Podesva, Robert & Sharma, Devyani. 2013. *Research methods in linguistics*. Cambridge University Press.

Price, Patti & Ostendorf, Mari & Shattuck-Hufnagel, Stefanie & Fong, Cynthia. 1991. The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America* 90(6). 2956–2970. doi:10.1121/1.401770

Roach, Peter. 1994. Conversion between prosodic transcription systems: "standard British" and ToBI. *Speech Communication* 15(1-2). 91–99. doi:10.1016/0167-6393(94)90044-2

't Hart, Johan & Collier, René & Cohen, Antonie. 1990. *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. New York: Cambridge University Press.

Taylor, Paul. 1998. The tilt intonation model. In *The 5th international conference on spoken language processing, incorporating the 7th australian international speech science and technology conference, sydney convention centre, sydney, australia, 30th november - 4th december 1998*. `http://www.isca-speech.org/archive/icslp_1998/i98_0827.html`.

Trager, George L. & Smith, Henry Lee. 1951. An outline of English structure. In *Studies in linguistics*, vol. 3 (Occasional Papers), Norman, OK: Battenburg Press.

Veilleux, Nanette & Shattuck-Hufnagel, Stefanie & Brugos, Alejna. 2006. 6.911 transcribing prosodic structure of spoken utterances with ToBI. MIT OpenCourseWare, `https://ocw.mit.edu`.

Wagner, Petra & Origlia, Antonio & Avezani, Cinzia & Christodoulides, George & Cutugno, Francesco & D'Imperio, Mariapaola & Mancebo, David Escudero & Fivela, Barbara Gili & Lacheret, Anne & Ludusan, Bogdan & Moniz, Helena & Chasaide, Ailbhe Ní & Niebuhr, Oliver & Rousier-Vercruyssen, Lucie & Simon, Anne-Catherine & Šimko, Juraj & Tesser, Fabio & Vainio, Martti. 2015. Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence. In Scottish Consortium for ICPhS 2015, The (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow: The University of Glasgow.

Welby, Pauline. 2006. French intonational structure: Evidence from tonal alignment. *Journal of Phonetics* 34. 343—371.

Xu, Yi. 2012. Focus and form in speech prosody–lessons from experimental research and potential implications for teaching. In Romero-Trillo, Jesús (ed.), *Pragmatics, prosody and English language teaching*, 61–76. Springer.

Yuan, Jiahong & Liberman, Mark. 2008. Speaker identification on the scotus corpus. Proceedings of Acoustics '08.