Prosodic Theory and Practice

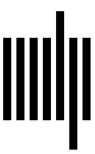
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel

Citation:

Prosodic Theory and Practice
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel
DOI: 10.7551/mitpress/10413.001.0001
ISBN (electronic): 9780262543194
Publisher: The MIT Press
Published: 2022

OA Funding Provided By:

OA Funding from MIT Press Direct to Open



The MIT Press

10 The PaIntE Model of Intonation

Antje Schweitzer, Bernd Möbius, Gregor Möhler, and Grzegorz Dogil

10.1 Introduction

The PaIntE model (Möhler and Conkie 1998; Möhler 2001) was originally developed for F0 modeling in text-to-speech (TTS) synthesis. Its purpose was to generate F0 contours that are as close to natural, human-produced F0 contours as possible. We will show in this chapter that it can also be used for more general research on intonation. The PaIntE model assumes that only the F0 contour in the vicinity of so-called intonation events contributes to the intonational meaning of an utterance, whereas the stretches in between these events arise from interpolation and do not affect the overall meaning. This view is manifested in the model's name: parameterized intonation events, or PaIntE for short.

PaIntE can be classified as a sequential model of intonation, in that it composes the F0 contour from a sequence of local contours, each associated with some kind of meaningful tonal event, and that these events or local contours do not interact or affect each other. PaInte shares this view with the well-known, phonologically motivated, tone sequence model (TSM; Pierrehumbert 1980), for instance. The sequential approach is also popular in speech synthesis, with the Tilt model (Taylor 1998) as probably the most widespread example.

While the PaIntE model has long been successfully used to generate perceptually appropriate F0 contours in speech synthesis, we will show in this chapter that PaIntE is also well suited for intonation research beyond speech synthesis.

To this end, we will introduce the PaIntE model in some detail in section 10.2, focusing on several aspects. First, we will show that PaIntE shares assumptions with autosegmental models of intonation (section 10.2.1). Second, in contrast to many other models, which aim to identify typical shapes that correspond to intonation categories, PaIntE assumes that several dimensions contribute to the meaning of tonal events and quantifies these dimensions by continuous parameters. The overall shape of the event is then determined by these parameters, as illustrated in some detail in section 10.2.2. Next, we provide more information on how the PaIntE parameters can be derived from a database (section 10.2.3) or how they can be predicted for speech synthesis (section 10.2.4).

In section 10.3, we will illustrate that the PaIntE intonation events can be related to categories posited by autosegmental approaches to intonation. Also, as we will show in section 10.4, PaIntE can be used for answering typical questions in intonation research. To this end, we will discuss some recent studies that have used the PaIntE model to investigate intonation, some from an autosegmental perspective, and some from an exemplar-theoretic perspective. Thus, sections 10.3 and 10.4 show that PaIntE

is compatible with an autosegmental approach to intonation, but it can also serve an exemplar-theoretic approach. Its flexibility lies in that, on the one hand, it can take autosegmental categories into account, and that the PaIntE parameters can be linked to these categories. On the other hand, however, it does not necessarily assume that such categories exist. We discuss the advantages of this property of PaIntE in section 10.5 and offer some conclusions in section 10.6.

10.2 The PaIntE Model and Its Parameters

To motivate the requirements and objectives of an F0 model in the context of speech synthesis, we briefly sketch the role of such a model in the TTS synthesis process. We then consider the commonalities between intonation models for synthesis and more general models of intonation from a theoretical point of view, before turning to the specific implementation of PaIntE in terms of its parameters and how these parameters can be extracted for analysis, or predicted for synthesis.

10.2.1 Intonation Models and Speech Synthesis

Traditional concatenative TTS systems generate speech starting out from a given, text-only specification of the utterance to be synthesized. This specification is passed through a pipeline of mostly independent modules, each of which incrementally adds linguistic, phonological, and phonetic information. Toward the end of this process, an F0 model adds concrete F0 values to the specification, and at the end of the pipeline, synthesized speech is generated by concatenating speech segments from a recorded database and, if necessary, manipulating these segments to match the F0 values that the F0 model has predicted.

What F0 models for synthesis and more general intonation models have in common, at the very least, is that they are interested in relating aspects of meaning to tonal contours. For instance, a TTS system might want to relate sentence-internal major syntactic phrase boundaries to rising intonation contours and sentence-final syntactic phrase boundaries to falling contours. Or it may relate exponents of information structure to contours conveying the intended meaning.

Many concatenative TTS systems treat the problem of predicting F0 contours following what may be called a phonological approach. According to Ladd (1996, 11), a phonological model of intonation has to minimally consist of two ingredients: first, a finite set of intonation categories, and second, a mapping from these categories to continuous acoustic parameters. In this vein, many TTS systems take into account linguistic properties inferred from the text to first predict the occurrence of a finite set of intonation categories such as pitch accents or boundary tones and then generate concrete F0 values in a second step. In fact, in the context of TTS, the term *F0 model* often refers just to this second step, that is, the generation of an F0 contour given a specification that already contains the desired location of some kind of intonation categories in the utterance.

In the case of a phonological approach to the TTS problem as just outlined, an additional commonality between F0 models for synthesis and more general intonation models is that both types of models have to address two issues: (i) identifying the relevant categories and (ii) specifying how these categories are implemented phonetically in terms of F0 (and, realistically, in terms of other prosodic cues such as duration).

A TTS system that follows such a phonological approach has the additional objective of specifying a mapping from linguistic properties to intonation categories. This is

not necessarily an objective of a more general model of intonation. However, to establish a distinction between two intonation categories, even a more general model would have to show that exchanging the two categories in some utterance context changes the meaning of the utterance.[1] Therefore, a more general intonation model cannot be entirely silent regarding the relation between meaning and intonation categories. The exact nature of this relation is still an open issue. For instance, there is consensus that in West Germanic languages such as English and German, information structure affects pitch accent placement (Terken and Hirschberg 1994; Féry and Kügler 2008) and even the type of pitch accent (Pierrehumbert and Hirschberg 1990; Chen, den Os, and de Ruiter 2007). However, to capture this impact, researchers have to refer to fine distinctions in information status that naive speakers are probably not aware of. Pierrehumbert and Hirschberg (1990), for example, elaborate differences among five meanings, termed *new*, *addition of new value*, *accessible*, *modification of given*, and *given* by Baumann, Röhr, and Grice (2015), that are claimed to give rise to different pitch accents. However, Baumann and Grice (2006) show that for German, a more fine-grained notion of accessibility is needed because accent types differ depending on the way in which the accessible information can be inferred from the text. Similarly, Baumann, Röhr, and Grice (2015) differentiate ten classes of information status that differ in accentedness and type. A crucial difference, then, is that a TTS system can rely only on more coarse-grained meaning that can be estimated from raw text, namely, text without annotation or markup, because this is what serves as input to a TTS system. Even worse, the TTS system has to expect that the estimated meaning may be incorrect at times.

Given these difficulties, most TTS systems treat the mapping from linguistic properties to intonation categories as a separate task, which is addressed as part of the linguistic analysis of the text to be synthesized (Sproat 1998; Taylor 2009). Then the task of the F0 model is "only" to map from a specification of the utterance, which already includes intonation categories, to concrete intonation contours. In other terms, the task of the F0 model in synthesis is to provide the phonetic implementation of phonological categories that have been determined in a preceding linguistic analysis step. This separate treatment is fostered by the dissemination of speech corpora with manually annotated tones and break indices (ToBI) labels (e.g., Ostendorf, Price, and Shattuck-Hufnagel 1996; Rapp 1998; Calhoun et al. 2010; Eckart, Riester, and Schweitzer 2012), which conveniently serve as training and testing data for this second step in F0 modeling.

Arguably, the separate treatment may also reflect a split in research approaches between studies that relate meaning to intonation categories (e.g., Beckman 1996; Büring 1997; Féry 1993; Pierrehumbert and Hirschberg 1990; among many others) and studies that investigate phonetic detail in the implementation of F0 contours (e.g., Pierrehumbert 1981; Kohler 1990; Ladd, Mennen, and Schepman 2000; van Santen and Möbius 2000; among many others).

The PaIntE model follows the practice of separating the prediction of categories from the actual F0 modeling. It does not necessarily state what the exact nature of the categories is. All that is said is that they are "intonation events," and the core task of the PaIntE model then is to generate concrete F0 contours that implement these intonation events. However, PaIntE acknowledges that ToBI categories are an obvious and convenient choice in that respect and provides means to take ToBI categories as the relevant intonation events for which local F0 contours have to be generated, by way of a configuration parameter. Before we go into detail on how the PaIntE parameters of a given contour can be extracted, and how the parameters can be predicted in speech synthesis, we first discuss how they determine the concrete F0 shapes.

10.2.2 The PaIntE Parameters

All intonation models dedicated to F0 modeling for speech synthesis parameterize the F0 contour in some way. In the case of PaIntE, the shape of the F0 contour around local intonation events is captured by six linguistically motivated parameters. Together they determine the F0 contour in a window of up to three syllables centered on the event. In the original implementation, the events were taken to be pitch accents and boundary tones posited by a German GToBI variant (Mayer 1995), that is, the PaIntE parameters served to specify the exact shape of the F0 contour on and around pitch accents and boundary tones. The global contour then arises by interpolation between these events.

Mathematically PaIntE employs a function of time, with $f(x)$ giving the F0 values at time x. It is defined as follows:

$$f(x) = d - \frac{c_1}{1 + e^{-a_1(b-x)+\gamma}} - \frac{c_2}{1 + e^{-a_2(x-b)+\gamma}} \tag{10.1}$$

This function yields a peak shape (figure 10.1), where the first term, the d constant, gives the upper bound. We will see that d can be interpreted as *peak height* parameter. From this d constant, two sigmoids are subtracted—the second and third terms in the equation. The first of these two sigmoids alone would result in a falling shape, starting at c_1 in negative infinity ($\lim_{x \to -\infty} = c_1$) and ending at 0 in infinity ($\lim_{x \to \infty} = 0$). The most pronounced part of this fall ends approximately at the value for parameter b. Because this sigmoid is subtracted from the d constant, this effectively yields a rise toward d, that is, toward the peak height parameter. The amplitude of this rise is c_1, and the pronounced part of the rise ends approximately at the value for parameter b. In the same way, subtracting the second, originally rising, sigmoid adds a fall component to this rising shape. The pronounced part of the fall starts close to parameter b, that is, approximately at the point where the first sigmoid levels off. Thus, we get a pronounced peak with the peak location affected by the b parameter; in other words, b can be interpreted as the *peak alignment* parameter. As parameter d is the upper bound for rise and fall, and thus the upper bound for the peak, this parameter corresponds to peak height. The amplitudes of the rising and falling parts are determined by parameters c_1 (rise amplitude) and c_2 (fall amplitude), and their steepness by parameters a_1 (steepness of rise) and a_2 (steepness of fall).

PaIntE provides several methods to normalize the time axis. In the standard variant, which is called *sylnorm* normalization, the time axis inside the approximation window is normalized such that syllable boundaries occur at integer values, with the syllable related to the intonation event beginning at 0 and ending at 1. In the sylnorm case, b determines the temporal alignment of the peak in terms of relative position within the syllables in the approximation window. A hypothetical example peak contour for a syllable associated with a pitch accent, using sylnorm normalization, is given in figure 10.1. The PaIntE function as specified in equation 10.1 is indicated by the solid line. Syllable boundaries are indicated by vertical lines, the syllables themselves are indicated by σ symbols, and the pitch-accented syllable is marked as σ*. The location of the b parameter (peak alignment) is marked by the bold vertical line; parameters c_1 (rise amplitude) and c_2 (fall amplitude) are indicated by the arrows, and the d parameter (peak height) by the tick at the y-axis. Parameters a_1 (steepness of rise) and a_2 (steepness of fall) cannot be read off the graphical representation in the same way as the other parameters, but they are hinted at in figure 10.1.

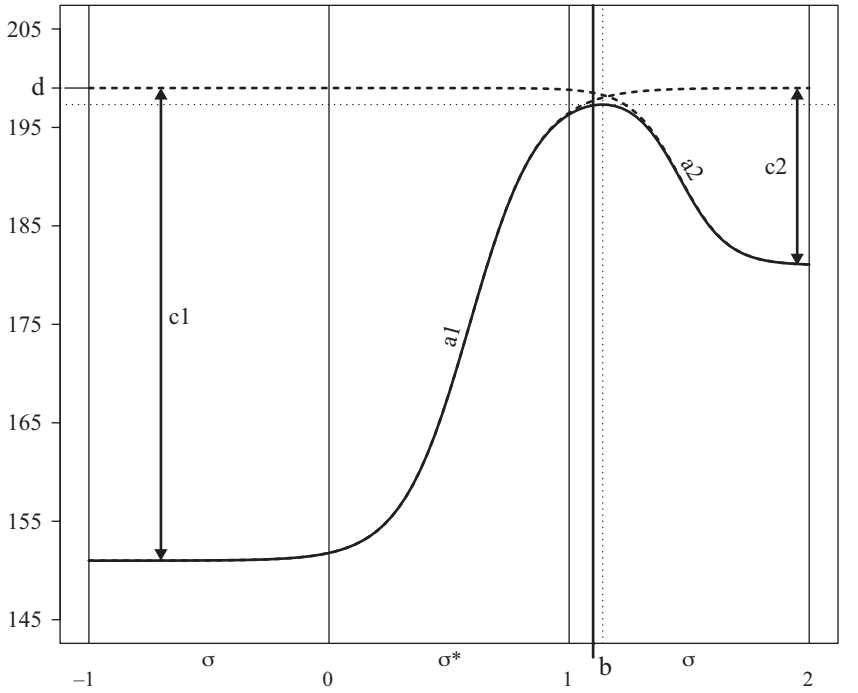


Figure 10.1
Example PaIntE contour in a window of three syllables around a pitch-accented syllable (σ*). See the text for details.

The shape with the pronounced peak is not prototypical for all syllables that are associated with some intonation event. Often, we observe just a falling or just a rising contour, without a clear peak. To accommodate such cases when parameterizing existing F0 contours, the PaIntE model first tries to detect a peak in the three-syllable window. It uses the PaIntE function as specified in equation (10.1) only in case there is a peak. If no peak is detected, the function is used with only the first sigmoid for rising contours or only the second sigmoid in case of falling contours, yielding the following two functions:

$$f_{rise}(x) = d - \frac{c_1}{1 + e^{-a_1(b-x)+\gamma}} \tag{10.2}$$

$$f_{fall}(x) = d - \frac{c_2}{1 + e^{-a_2(x-b)+\gamma}} \tag{10.3}$$

We have indicated these alternative functions by dashed lines in figure 10.1. They are partly hidden by the PaIntE function (solid line), but it can be seen that the *b* parameter, indicated by the bold vertical line, does indeed occur at the point where the dashed line for the rising sigmoid starts to level off, and where the fall in the dashed line for the falling sigmoid is about to become more pronounced. It is thus a reasonable estimate of the end of the rise, given that, mathematically, the rise does not end at all

because it never reaches *d* before infinity. Similarly, the fall effectively starts in (negative) infinity, not somewhere in the window depicted here. The same argument can be made for parameter *d*; it is only an estimate for the height of F0 at the end of the rise, or the beginning of the fall, respectively, and is never really reached. However, it can be seen in figure 10.1 that for the two example sigmoids, the difference between *d* and the function values at the edges of the three-syllable window is not detectable by eye; the dashed line for the falling sigmoid seems to reach the *y*-axis exactly where the tick for *d* is located.

It should also be noted here that in the case of the "full" PaIntE function with the peak as given in equation (10.1), the *b* parameter is again only an approximation of the temporal alignment of the peak in the syllable structure. It can be seen in figure 10.1 that the peak's exact temporal location, as indicated by the dotted line, is actually a small distance to the right of the *b* parameter itself, which is indicated by the bold vertical line. The exact displacement depends on the values of the a_1, a_2, c_1, and c_2 parameters. In this specific example, for instance, the true peak is 0.04 units to the right of the *b* value, that is, if the last syllable in this example were two hundred milliseconds long, *b* as an approximation of the temporal alignment of the peak would be off by eight milliseconds. If a better estimate of the peak alignment is desired, one can resort to sampling the curve specified by the PaIntE function and finding the point where the maximum sample occurs.[2]

To get an impression of the accuracy of *b* for approximating the location of the peak covering a representative number of contexts, we have estimated both *b* and the true temporal alignment of the peak for approximately seventeen thousand syllables in a database of two hours of speech that had been approximated using the full PaIntE function in equation (10.1). The mean absolute error in syllable units was approximately 0.052 and the median 0.039. In absolute time, this corresponded to a mean absolute error of approximately ten milliseconds and a median of approximately eight milliseconds. Because the time resolution in deriving F0 from the speech signal is usually in this order, we consider the approximation to be exact enough. However, if necessary, it is easy to estimate the true location of the peak given the six PaIntE parameters as suggested here.

The same point can be made for the *d* parameter (peak height). In the example in figure 10.1, the *d* parameter is 199 hertz, while the true peak height is at approximately 197.3 hertz. For the seventeen thousand syllables from our database, the mean absolute error was approximately 1.380 hertz, and the median absolute error was approximately 1.167 hertz. This is on the order of the just noticeable difference of approximately one hertz in human perception of complex tones at pitch levels below five hundred hertz (Kollmeier, Brand, and Meyer 2008). This leads us to conclude that the approximation is accurate enough for almost all purposes. If a higher accuracy is needed, we again recommend estimating the true peak height numerically by sampling.

So far we have discussed only the application of PaIntE using the sylnorm normalization; however, PaIntE also provides an alternative called *anchor* normalization. In this case, each syllable is split into three parts representing the (unvoiced) onset of the syllable; its sonorant nucleus, which is defined as containing the nucleus and possibly preceding voiced consonants in the onset; and, finally, the coda. This normalization is motivated by findings that indicate that timing in F0 movements is relative to syllable structure (House 1996, 1997; van Santen and Möbius 2000). Using anchor normalization, each syllable in the approximation window is again normalized to length one with the same values for syllable boundaries as in the sylnorm case. Syllable-internally,

the unvoiced onset is adjusted linearly to a length of 30 percent of the syllable duration. The sonorant nucleus then spans another 50 percent: it ranges from 30 to 80 percent of the syllable duration, and the coda finally goes from 80 to 100 percent of the syllable duration.

It should be noted that the approximation window may be shorter than three syllables depending on the context. This is the case if silent intervals intervene between the syllables. Also, as stated, PaIntE can be configured to take information regarding prosodic categories associated with the syllables into account. In this case, the window does not extend to neighboring syllables that also carry a pitch accent, or across syllable boundaries that are associated with a phrase boundary. Reducing the approximation window in these cases is motivated by the fact that speakers before boundaries compress tonal contours that in another context would extend into the following syllables (Mayer 1995; Grabe 1998; Jilka, Möhler, and Dogil 1999). Similarly, in cases where it is known that a syllable is associated with a pitch accent that exhibits a late peak (e.g., in an L*H accent as assumed by Mayer 1995), the approximation window does not contain the preceding syllable. However, it is possible to configure PaIntE to enforce the three-syllable window in all contexts, just as it is possible to parameterize every single syllable in cases where no prosodic annotation is available that indicates where intonation events are expected.

10.2.3 Extracting the PaIntE Parameters

Before the PaIntE parameters of a given F0 contour can be approximated, the raw F0 contour is smoothed to eliminate micro-prosodic effects and outliers. To this end, PaIntE uses the smooth_f0 algorithm based on the Edinburgh speech tools (Taylor et al. 1999). Smooth_f0 is a median smoother that interpolates across unvoiced regions, but not across silences.

In preparing the approximation, PaIntE first looks for an F0 peak in the smoothed contour in the middle of the approximation window, as well as local minima to the right and left of the maximum. The locations of the maximum and the minima, as well as the number of frames available, are used to determine which of three approximation methods is appropriate in that particular context:

- *Mean F0 approximation*. No PaIntE approximation takes place if there are fewer than two voiced frames for the current window or if the two minima are less than five frames apart. In these cases, PaIntE reverts to a simple approximation called *meanf0* by just determining the mean F0 value in that window as the *d* parameter (peak height); the five other PaIntE parameters are set to 0.
- *Single sigmoid approximation*. If either the left or the right minimum coincides with the maximum, that is, a rise or fall has been detected, but a clear peak has not, the PaIntE approximation is modified to leave out one of the two sigmoids, as described in section 10.2.2. Depending on which sigmoid is left out, this is called the *rise_sigmoid* or the *fall_sigmoid* method. In this case, the *a* parameter (steepness of rise or fall) of the missing sigmoid is set to –1, and its *c* parameter (rise or fall amplitude) is set to 0. The remaining parameters are determined by the single sigmoid approximation.
- *PaIntE approximation*. In the standard case, in which a peak has been detected, that is, neither minimum coincides with the maximum, the approximation is carried out using the PaIntE function as defined in equation (10.1). This is called the *pfun* method.

The approximation itself determines the PaIntE parameters using the appropriate functions, choosing the parameters so that the root mean squared error (RMSE) between actual F0 values and the corresponding values in the PaIntE function is minimized. Finding the optimal combination of parameters is an optimization problem, and PaIntE uses a conjugate gradient method to arrive at a local optimum.

10.2.4 Predicting the PaIntE Parameters

There are several ways in which the PaIntE parameters can be used to generate F0 contours in speech synthesis. As discussed, speech synthesis systems often approach predicting F0 contours as a two-step problem, in which first a set of intonation categories, for instance, the ToBI categories, is predicted from the text. As a result, one would have a specification of the utterance to be synthesized that already includes concrete ToBI categories. The task of F0 modeling can then be viewed as mapping from the ToBI category to the PaIntE parameters, taking linguistic and phonological context into account.

Typically, this mapping would be learned using machine learning techniques on a database that is annotated with all the context properties that will be available at synthesis time, and with the category labels. Most machine learning schemes predict only one parameter at a time; thus, one approach would be to train six models, each of which predicts one PaIntE parameter given the context. However, even if each predicted parameter may be plausible, this does not ensure that the combination of the six predicted parameters is plausible too. One way to avoid this problem is to first determine a finite number of "typical" combinations using clustering techniques. For instance, Möhler and Conkie (1998) used vector quantization, experimenting with between four and thirty-two clusters, whereas Möhler (2001) used up to sixty-four clusters. Then the actual F0 modeling consists of mapping from the given context to one of the clusters instead of to the continuous parameters, thereby turning the regression problem into a classification problem.

The idea of determining a number of typical phonetic implementations by clustering may lead to the question of whether the ToBI categories could be completely replaced by clusters found in this way. However, Möhler (2001) found that the results are better if the GToBI categories proposed by Mayer (1995) are taken into account, both in terms of RMSE between synthesized contour and original, and in terms of correlation between the two. This finding is just one indication that there is a correlation between ToBI-like categories and the PaIntE parameters. In the following section, we show that there is indeed a systematic relationship between these categories and the PaIntE parameters, as the parameters reflect properties that are related to the defining characteristics of the ToBI categories.

10.3 Relating PaIntE to Prosodic Categories

The PaIntE parameters can be related to established categories in a straightforward way. To demonstrate this, we will show here that the PaIntE parameters reflect the expected shape of contours associated with ToBI-style categories—in our case, the tonal categories assumed by the GToBI variant proposed by Mayer (1995). We refer to this variant as German ToBI (Stuttgart variant), or GToBI(S). To this end, we examine PaIntE parameters of pitch-accented syllables extracted from a large database of German read speech.

The database that we use for this purpose was recorded for unit selection speech synthesis (Barbisch et al. 2007) in the course of the SmartWeb project (Wahlster 2004). It was read by a professional male speaker of Standard German and contains typical,

isolated utterances of five different genres, usually consisting of one, or at most two, short sentences, corresponding to several prosodic phrases. All utterances were annotated on the segment, syllable, and word level, and prosodically labeled according to GToBI(S). Prosodic labeling for each utterance was carried out by one of three human labelers, supervised and instructed by the first author, in the process of building a database for unit selection speech synthesis. The database amounts to two hours of speech, containing seventy-two thousand segments, twenty-eight thousand syllables, and fourteen thousand words.

GToBI(S) assumes five basic types of pitch accents: L*H, H*L, L*HL, HH*L, and H*M, sometimes described as rise, fall, rise-fall, early peak, and stylized contour, respectively. Just like other ToBI approaches, it assumes that pitch accents are characterized by either a high (H) or a low (L) target associated with the accented syllable and indicates this association using the starred tone notation, H* or L*. It also assumes that the contour before and after the starred tones is determined by trailing and leading tones. In contrast to many other ToBI variants, the notation for these tones is without a + sign to separate trailing and leading tones from the starred tone, but this is a purely notational difference.

Another, less trivial, difference is that GToBI(S) allows tritonal accents: the L*HL accent has two trailing tones, H and L, and the HH*L accent has both a leading H and a trailing L tone. Also, it assumes that the L*H accent and the H*L accent have slightly less prominent allotonic variants, that is, alternative realizations that do not change the underlying meaning of the accent. Mayer (1995) suggests that they can be realized by just the starred tone on the accented syllable, and that the trailing tone can be split off and realized later (partial linking), or can even be omitted completely (complete linking). This results in two monotonal accents, H* and L*, that are interpreted as variants of the H*L and L*H accents, respectively, which differ only in perceived prominence, but not in meaning. Another important difference to the widespread GToBI labeling scheme proposed by Grice and colleagues (Grice and Baumann 2002; Grice, Baumann, and Benzmüller 2005) is that GToBI(S) does not distinguish between an L + H* and an L* + H pitch accent. GToBI(S) provides only the latter category, namely, L*H in Mayer's notation. Cases where other GToBI variants assume L + H* are accounted for in other ways, for instance, by assuming a monotonal H*, where the low pitch level just before the accented syllable is caused by other factors, for instance, by partial linking of a preceding accent.

Apart from this, the expected shape of the pitch contour for each accent is manifested in its notation, as in all ToBI dialects. The contour is expected to reach the target for the starred tone ideally in the middle of the accented syllable, the targets for leading tones on the preaccented syllable, and the targets for the trailing tones on the postaccented syllable or syllables.

Figure 10.2 shows parameterization results for nine pitch-accented syllables selected for illustration from the above-mentioned database. We find that the properties expected given the GToBI(S) categories are reflected in the concrete contours, although there are some differences in the detailed implementation. For instance, in the first accent, identified as an L*HL accent, the rise already starts at the beginning of the accented syllable, reaches the peak at the boundary to the following syllable, and falls within the postaccented syllable. It is thus realized in a more compressed way than expected given its description by Mayer (1995). In the middle panel of the first row, the peak in the H*L accent is at the boundary between the accented and the postaccented syllable, that is, later than the middle of the accented syllable. In the next three

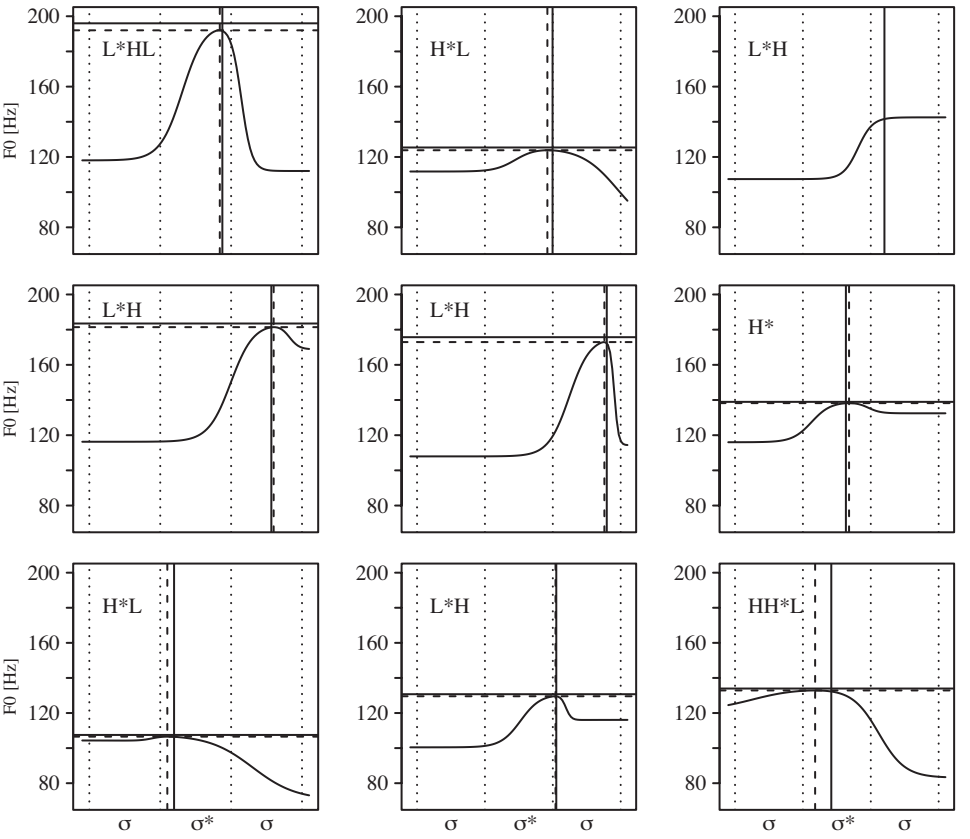


Figure 10.2
PaIntE contours for some GToBI(S) pitch accents. Dotted lines indicate syllable boundaries, σ symbols represent syllables, and the accented syllable is indicated by asterisks. Solid vertical lines indicate the peak alignment parameter b, solid horizontal lines indicate the peak height parameter d, and dashed lines indicate the "true" peak location as estimated using the sampling method outlined in the text. The type of pitch accent according to manual prosodic annotation is indicated within each panel.

accents, all L*H accents, the rise starts on the accented syllable and continues into the next syllable, reaching the peak early in this syllable (in the right panel in the first row), well within this syllable (in the left-most panel in the middle row), or late in this syllable (middle panel in the middle row). Note that in this last example, the contour exhibits a pronounced peak with an abrupt fall, which could be due to a reset to a low pitch level for the next target and thus does not necessarily reflect a property of the L*H accent in question. It should also be noted that the first of these three L*H accents does not exhibit a peak: it was parameterized using the function with only the rising sigmoid.

In the following panel (the right-most panel in the middle row), we can see a prototypical example of a monotonal H* accent, with a less pronounced, broad peak, and relatively low amplitudes, which corresponds well to its characterization as being less prominent than the bitonal H*L variant. Similarly, the three accents in the bottom row reflect the expected properties well: the H*L has a high target in the accented syllable, the L*H rises to a peak at the boundary of the next syllable, and in the HH*L, the

contour is already high throughout the preceding syllable and falls to a low level in the postaccented syllable.

To return once more to the question of how accurate the *b* and *d* parameters are, we have indicated the "true" values for peak height and peak alignment by dashed lines in all cases where the approximation was carried out using the full PaIntE function. It can be seen that sometimes there is a small gap between the estimated peak height (solid horizontal line) and the "true" peak height (dashed horizontal line), as evidenced in the top left panel and in the first two panels in the middle row. However, relative to the corresponding rise and fall amplitudes, the difference is small. Similarly, the "true" temporal alignment, indicated by the dashed vertical line, is occasionally at some distance to the *b* parameter as specified in the PaIntE function, indicated by the solid vertical line. This is most obvious in the top middle panel and in the bottom left panel, as well as in the bottom right panel. However, as can be seen in these examples, these larger differences always occur in cases with broader peaks, where one could argue that it is hard to tell where exactly the peak should lie anyway.

These examples demonstrate that it is possible to find the expected properties of pitch accents reflected in the PaIntE parameterization results. But does this also hold on a larger scale, across many examples? To investigate this question, we examine density plots of parameters *b* (peak alignment), c_1 (rise amplitude), c_2 (fall amplitude), and *d* (peak height), obtained from the above-mentioned database. Because GToBI does not make any predictions about the steepness of the contours, we will not discuss parameters a_1 (steepness of rise) and a_2 (steepness of fall) here. Because of limitations in space, we will address only the most frequent of the basic GToBI(S) accents: the L*H, H*L, and L*HL accents.[3]

Figure 10.3 shows density plots of the *b* parameter (*peak alignment*). They are based on parameterization results of the approximately 3,200 syllables with L*H accents, approximately 1,800 syllables with H*L accents, and approximately 270 syllables with L*HL accents in our database. Density plots show how likely a certain range of values is in the underlying data: peaks appear at values that are more likely to occur for the underlying sample, whereas valleys appear at values that are less likely to occur.

We again indicate syllable boundaries by vertical lines. Thus, the broad peak in the dashed line for H*L accents in the left panel, between the two vertical lines, indicates that H*L accents are most likely to have their peak in the middle of the accented syllable. For L*HL accents, which are indicated by the dot-dashed line, surprisingly, the peak is also on the accented syllable. From the description of L*HL accents given by Féry (1993, 94) and Mayer (1995), one would expect this peak to be on the postaccented syllable in many cases. Nevertheless, compared to H*L and H*, the peak for L*HL accents is shifted further toward the syllable boundary. It is also slightly narrower than the two peaks of the H*L and H* distributions, indicating less variation of *b* for L*HL accents.

Finally, the density for L*H accents (solid line) is bimodal: L*H accents are almost equally likely to have their peak either right before the syllable boundary, just as L*HL accents did, or in the later part of the postaccented syllable. One might interpret this as evidence for two distinct categories L + H* (with the peak on the accented syllable) and L* + H (with the peak on the postaccented syllable) as in the more widespread GToBI variant (Grice and Baumann 2002; Grice, Baumann, and Benzmüller 2005); however, the right panel in figure 10.3 shows that this bimodal distribution comes about because L*H is realized differently on word-final syllables than on nonfinal syllables.[4] Here, the dashed line represents L*H accents that occurred on word-final syllables, and the dot-dashed line represents L*H accents that occurred on word-internal syllables. Obviously, these two contexts cause the bimodal distribution: L*H accents on word-final syllables

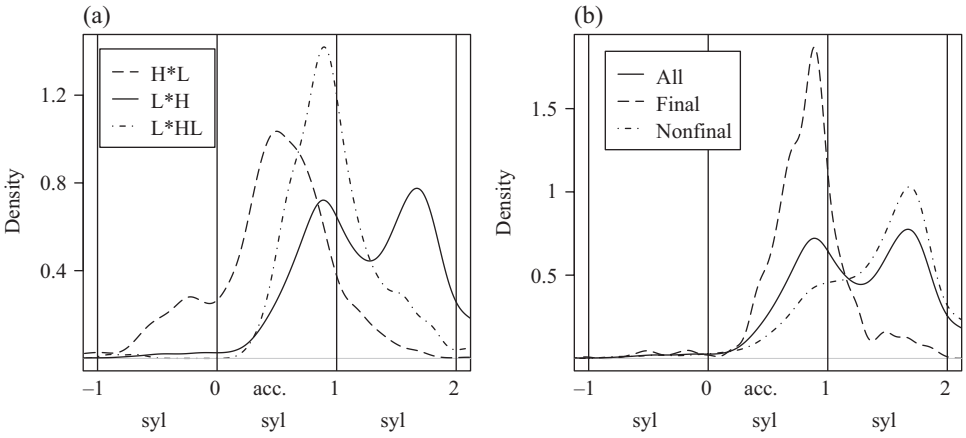


Figure 10.3
Density plots of the b parameter (peak alignment) for some GToBI(S) accents. (Left) H*L accents (dashed line) have their peak earlier in the accented syllable than L*HL accents (dot-dashed line). L*H accents (solid line) have their peak on either the accented or the postaccented syllable. (Right) The bimodal distribution for L*H accents (solid line, repeated from left panel, different scaling) obviously arises because L*H accents in word-final syllables (dashed line) have their peak on the accented syllable, while L*H accents in word-internal syllables (dot-dashed line) have their peak on the following syllable. acc, accented; syl, syllable.

almost always have their peak in the accented syllable, while word-internal L*H accents tend to have their peak on the postaccented syllable. In other words, the tonal movement on L*H accents usually does not cross word boundaries; instead, it is timed to occur earlier before word boundaries.

The example of the alignment parameter *b* shows that the PaIntE parameters not only capture well-known properties of the GToBI(S) accents, but that they can also serve to investigate context-dependent aspects of phonetic implementation, as in the case of word-internal versus word-final L*H accents.

Figure 10.4 gives density plots for parameter *d* (peak height). Values for *d* in hertz are indicated on the *x*-axis: values to the right indicate higher peaks. The figure thus shows that syllables associated with H*L accents (dashed line) are very likely to exhibit low values for *d*: the peak in the line indicates that values just below 120 hertz for peak height are most probable. Compared to L*H (solid line) and L*HL (dot-dashed line), the peaks of H*L accents are lower. This is due to the prevalence of nuclear, that is, phrase-final, H*L accents over prenuclear H*L accents: ninety-three percent of the H*L accents in the database are nuclear accents. For nuclear accents, lower peaks must be expected because of F0 declination, that is, the global trend of F0 to decline over the course of the utterance (e.g., Cooper and Sorensen 1981; Gussenhoven and Rietveld 1988; Pierrehumbert 1979). Indeed, the density plot for non-nuclear H*L accents (not depicted here) is shifted to the right and is broader, similar to the distribution for L*H accents. Peaks of L*HL accents (dot-dashed line) are high even though they are usually nuclear accents in our data (84 percent). The distribution for L*H accents (solid line) is similar to that of L*HL accents (dot-dashed line), although there is again more variation for L*H accents.

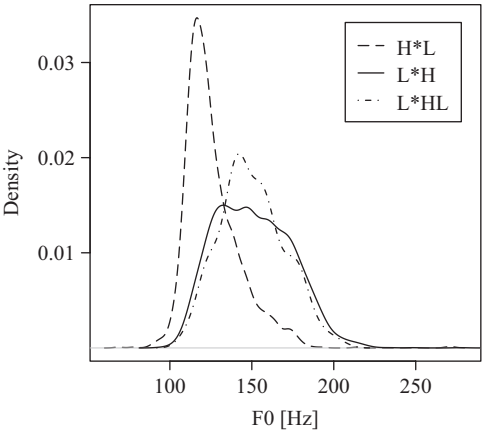


Figure 10.4
Density plots of the *d* parameter (peak height) for the most frequent accents and for unaccented syllables. H*L accents (dashed line) exhibit the lowest values for *d*; L*HL accents (dot-dashed line) and L*H accents (solid line) are characterized by higher and more variable values for *d*.

The last two PaIntE parameters to be discussed here, c_1 and c_2, determine the amplitude of the rise toward the peak (c_1) and the amplitude of the fall after the peak (c_2) in hertz. Figure 10.5 shows the distributions of c_1 (left panel) and c_2 (right panel) for different accent types. Looking at H*L accents first, which are indicated by the dashed line, there is little surprise. It is obvious that they tend to have low values of c_1 (rise amplitude) but higher values of c_2 (fall amplitude): their c_1 distribution shows a pronounced peak for c_1 values of around zero to ten hertz, and although the distribution extends to the right with values of c_1 up to sixty to eighty hertz, the higher values are much less likely. Their c_2 distribution, however, shows a clear dominance of moderately high c_2 values with values of around zero being rather improbable. There is a broad peak between twenty and twenty hertz, indicating that these are typical values of c_2 for H*L accents. In short, H*L accents have small rise amplitudes but higher fall amplitudes, as expected for falling accents. L*H accents (solid line) show just the opposite behavior: their c_1 values are typically between twenty and sixty Hz, while their c_2 values tend to be close to zero, as one would expect for rising accents. The distributions for L*HL accents are given by the dot-dashed lines. They exhibit higher values for both c_1 and c_2, reflecting their characterization as rise-fall accents.

Thus, we have shown for the PaIntE parameters *b* (peak alignment), *d* (peak height), c_1 (rise amplitude), and c_2 (fall amplitude) that their distributions differ depending on which GToBI(S) pitch accent they are associated with, and that the PaIntE model captures the tonal characteristics of the pitch accents well. As a direct consequence of its versatility and accuracy, as well as its linguistic underpinnings, the PaIntE model has recently been employed in a number of intonation studies, which we detail next.

10.4 PaIntE in Intonation Research

In this section, we present several case studies to illustrate PaIntE's potential for intonation research. Moreover, we demonstrate that intonation modeling by means of

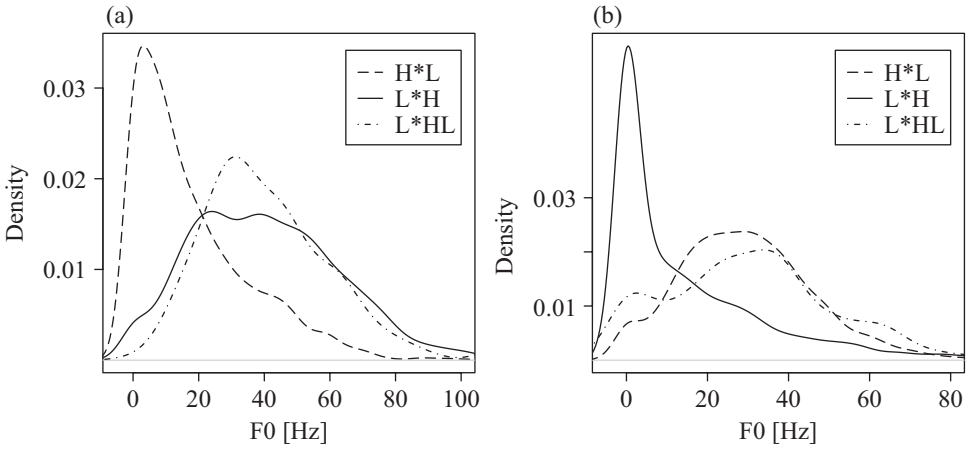


Figure 10.5
Density plots of the c_1 parameter (rise amplitude, left panel) and the c_2 parameter (fall amplitude, right panel) for the most frequent accents. H*L accents (dashed line) usually exhibit low c_1 and high c_2 values; L*H accents (solid line) exhibit low c_2 and high c_1 values. L*HL accents (dot-dashed line) are characterized by high values of both c_1 and c_2.

PaIntE can subserve both the autosegmental-metrical approach (section 10.4.1) and an exemplar-theoretic approach to intonation research (section 10.4.2).

10.4.1 Autosegmental Case Studies

We have shown that general rules concerning the phonetic implementation of intonation categories can be detected in the distributions of the PaIntE parameters as in the case of peak alignment in L*H accents discussed in section 10.3. This methodology can also be employed to test more general hypotheses regarding the implementation of F0 contours. For instance, Dogil and Schweitzer (2011) investigated the alignment of F0 peaks in several German and English databases in this way. Their hypothesis was that there is a quantal effect in peak alignment that causes speakers to place F0 peaks either before or after syllable onsets, but not within onsets. This hypothesis can be motivated by House's (1996) model of tonal perception, which claims that tonal contours within onsets are perceived differently from contours in the nucleus or coda, or by the observation that syllable onsets are considered "weightless" (Goedemans 1998).

To investigate whether peaks in F0 systematically avoid syllable onsets, Dogil and Schweitzer (2011) modified the PaIntE anchor normalization method described in section 10.2.2. Originally, using this normalization, in each syllable, the unvoiced onset is adjusted linearly to take up the first 30 percent of the syllable duration, the voiced onset together with the nucleus to range from 30 to 80 percent, and the coda to span the remaining 20 percent of the syllable duration. In the modified version, voiced and unvoiced onset consonants were treated the same, that is, the onset, regardless of whether it was voiced or unvoiced, was always mapped to the first 30 percent of the syllable duration.

Using this modified normalization, Dogil and Schweitzer (2011) extracted PaIntE parameters from the above-mentioned unit selection database, as well as from a very similar database of a female speaker. Both databases had been manually prosodically

labeled according to GToBI(S). The density plots for all pitch accents in these databases exhibited valleys within syllable onsets in their distributions for the *b* parameter (peak alignment), that is, both speakers avoided placing the peak within onsets. The same procedure was applied to a part of the Switchboard corpus of telephone conversations between nonprofessional speakers (Godfrey, Holliman, and McDaniel 1992), which was annotated for accent location (Calhoun et al. 2010), and to German audiobook recordings from the Librivox project,[5] for which no prosodic annotations were available. The valleys were also present in the onsets when looking at the distributions of just the accented syllables from Switchboard, and when looking at the distributions of all syllables in the audiobook corpus, irrespective of whether they were accented.

Investigating PaIntE parameter distributions is not the only way to carry out intonation research using the PaIntE model. It is also possible to investigate the parameters in a more direct way, for instance, by fitting linear mixed models to find which factors affect the PaIntE parameters. For instance, Kelly and Schweitzer (2015) used PaIntE to investigate lexical accents in Trøndersk, a dialect spoken around Trondheim in central Norway. Norwegian distinguishes two lexical accents, named accent 1 and accent 2, respectively. Previous research had found that in Trøndersk, the two accents have a similar shape with a high target followed by a low target, but that they differ in the alignment of tones with the segmental string, with a later timing for accent 2 (Kristoffersen 2006). Also, accent 2 had been shown to have a higher F0 minimum than accent 1 (Kelly and Smiljanić 2014).

Using PaIntE, Kelly and Schweitzer could confirm the previous findings on the later timing of the peak in accent 2: they found that a linear mixed model predicting syllable-normalized parameter *b* (peak alignment) with accent type as a fixed effect was significantly better than the corresponding model without accent type, that is, peak alignment depends on accent type. The study also provided new findings, that is, that accent 2 has a higher F0 maximum than accent 1, and that the amplitude of the fall is smaller in accent 2, again by showing that parameters *d* (peak height) and c_2 (fall amplitude) depend on accent type.

Kelly and Schweitzer's study is, to our knowledge, the first to use PaIntE to investigate lexical accents. The results demonstrate that the PaIntE parameters can be used to assess aspects of phonetic implementation of lexical accents yielding observations comparable to "classical" implementation studies that measure F0 maxima, minima, or turning points. The advantage of the PaIntE model is that these measurements can be derived automatically. This facilitates the investigation of intonation using data on a much larger scale.

10.4.2 PaIntE in Exemplar-Theoretic Approaches

We have demonstrated in section 10.3 that the PaIntE parameters are compatible with an autosegmental view of intonation in that they can serve to specify, or investigate, detailed context-dependent phonetic implementations of ToBI-like categories. This is in fact what PaIntE was designed for originally. Here we will show that PaIntE can also subserve an exemplar-theoretic account of intonation. We briefly introduce the ideas behind examplar theory, and then we give examples of research that has used PaIntE to this end.

In recent years, exemplar theory has gained increasing attention, especially in the segmental domain. The key idea in exemplar theory as applied to speech (e.g., Lacerda 1995; Goldinger 1996, 1997, 1998; Johnson 1997; Pierrehumbert 2001, 2003) is that speakers have access to memory traces (*exemplars*) of previously perceived instances of

speech in which almost full phonetic detail is retained. Categorizing new instances in speech perception is based on the stored exemplars and their categories (Lacerda 1995; Johnson 1997; Pierrehumbert 2001, 2003); in speech production, production targets are derived from stored exemplars (Pierrehumbert 2001, 2003).

Under an exemplar-theoretic account, phonetic categories are instantiated by accumulations of similar exemplars in memory. It is sometimes claimed that exemplar models negate abstraction in speech production and perception, but this is not the case. The difference from abstractionist models is that exemplar models assume that abstraction arises as a consequence of generalizing over a large set of exemplars (Pierrehumbert 2003). The aggregation of many exemplars with fine phonetic detail implicitly yields a more abstract linguistic concept with all the properties that the exemplars have in common, leaving all the details in which they vary underspecified. Often it is even explicitly assumed that the exemplars contain category labels (Johnson 1997; Pierrehumbert 2001; Walsh et al. 2010; Wade et al. 2010).

Few studies have looked at prosody in an exemplar-theoretic framework. However, it was shown already in one of the first exemplar-theoretic studies (Goldinger 1997) that exemplars seem to retain prosodic detail, in addition to segmental phonetic properties. In shadowing experiments, subjects tended to adapt their pitch from their baseline pitch toward the pitch of the stimulus token, and to match the durations of their productions to the stimulus token. The effect was stronger for low-frequency words. This indicates that pitch and duration are stored with the word exemplars and that these properties are retained in production.

Building on this, Schweitzer (2011) suggested that even more fine-grained prosodic properties, such as peak height, peak alignment, or rise and fall amplitudes, as quantified by the PaIntE parameters, might be stored in memory. In this vein, Calhoun and Schweitzer (2012) proposed that words and short phrases in American English are stored with their intonation contours and that discourse meanings of highly frequent word-contour pairings can spread by analogy to less frequent pairings. To substantiate this claim, they used PaIntE to parameterize the contours and calculated duration *z*-scores for the segments. Representing the contours by attributes derived from these parameters, they identified fifteen "typical" contours using clustering techniques. They found that certain words and contours form collocations, that is, they appeared together more often than would be expected based on their individual frequencies, supporting the hypothesis that words are stored together with their contours. In a perception experiment, they then confirmed that the discourse meanings of the most frequent pairings spread to other word-contour pairings, which constitutes evidence that the contours were indeed lexicalized.

Further evidence for exemplar storage of prosodic properties comes from a series of three experiments (Schweitzer et al. 2015) demonstrating that phonetic implementation of pitch accents, again in terms of PaIntE parameters, is subject to frequency of occurrence of the linguistic context. We address only the first of these three experiments in more detail here, as it investigates accent implementation in terms of PaIntE parameters directly and thus can serve to illustrate how the PaIntE parameters could be interpreted as dimensions in storing intonation contours. The other two experiments also utilize the PaIntE parameters; however, they are used to assess frequency effects on the similarity or dissimilarity of accent shapes.

The experiment described by Schweitzer et al. (2015) uses a database that was manually annotated for GToBI(S) pitch accents (Mayer 1995). Using generalized linear mixed models, the authors show that accent range in L*H and H*L accents, as quantified by

PaIntE parameters c_1 and c_2, respectively, is significantly affected by the frequency with which the accent and the specific word co-occur. Traditional autosegmental models of intonation, which assume that intonation is postlexical, cannot easily account for such frequency effects, while exemplar models offer a parsimonious account. It is assumed that in production, a number of exemplars that match the required target best are activated and that speakers average over these exemplars, or randomly sample from them, to arrive at a concrete production target. Thus, if a pitch-accented word is to be produced, and if sufficient pitch-accented instances of this word are stored in the speaker's memory, the derived target will match those exemplars and is thus expected to exhibit an F0 amplitude that is appropriate for pitch-accented words. If, on the other hand, however, only a few pitch-accented exemplars of this word are stored, other, nonaccented exemplars will contribute in deriving the production target, leading to a reduced F0 amplitude.

In summary, we would like to argue here that in an exemplar-theoretic account of intonation, the detailed intonational properties that are assumed to be stored with each exemplar can be captured by the PaIntE parameters. We do not necessarily advocate an exemplar-theoretic approach to intonation, but we would like to note that given the problems with labeler consistency and human labeling time, which will be discussed in the following section, an exemplar account does have a certain appeal. However, it should be noted that at least in the second study discussed here (Schweitzer et al. 2015), it is still assumed that exemplars might be labeled with concrete intonation category labels, that is, it does not make such categories obsolete.

10.5 Discussion

In this section, we discuss several theoretical and practical problems of intonation modeling arising from the assumption that the intonation structure of utterances can be described in terms of a linear sequence of intonation events that represent intonational categories. We then move on to discuss the characteristics of PaIntE that allow for a mapping of F0 contours to established intonational categories, but also for analyzing and generating F0 contours in a scenario in which one prefers to remain agnostic with respect to the validity of such categories.

Autosegmental models of intonation aim at establishing a set of intonation categories that, analogously to phonemes in the segmental domain, serve to distinguish meaning. This idea has driven most intonation research in the past fifty years or so (e.g., Goldsmith 1976; Bruce 1977; Gussenhoven 1984; Ladd 1996), with the TSM (Pierrehumbert 1980) and its extension to the ToBI labeling system for American English (Jun, chapter 4, this volume) as one of its most prominent and probably most widely accepted approaches. However, the categories proposed by these models are far from being as established as their segmental counterparts. Even models that do agree on the autosegmental approach differ in the specific inventory of categories that they suggest. In the case of American English, Dilley and Brown (2005), for instance, propose a set of categories for American English that differs from that of the TSM, or ToBI. Similarly, for German, a number of models have been proposed in the autosegmental tradition, all of which assume different category inventories (Kohler 1991; Féry 1993; Mayer 1995; Grice, Baumann, and Benzmüller 2005; Peters 2014). The ToDI transcription system for Dutch intonation (Gussenhoven 2005) focuses on the transcription of tones and limits the number of boundary categories to two phrase boundaries, namely utterance and intonation, without accounting for different strengths of boundaries (unlike ToBI).

Setting aside the problem of agreeing on one authoritative set of categories, a further problem is that even if the categories are taken as given, it is not easy to unambiguously identify these categories in speech data, as evidenced by the moderate consistency with which human labelers can identify them. For instance, regarding labeler consistency for ToBI pitch accents in spontaneous data from the Switchboard corpus (Godfrey, Holliman, and McDaniel 1992), Yoon et al. (2004) report a kappa coefficient of $\kappa \approx 0.51$ for interannotator agreement on the type and presence of pitch accents.[6] However, these moderate consistencies are achieved only when collapsing ToBI pitch accents into two broad categories H* and L*, plus a class X* for uncertain cases; the consistency for the original ToBI inventory must be expected to be even lower. Another study (Syrdal and McGory 2000) on read speech using the original ToBI inventory reports more promising values of $\kappa \approx 0.67$ for ToBI pitch accents in a male corpus and of $\kappa \approx 0.69$ in a female corpus, indicating substantial, but far from perfect,[7] agreement. These corpora, however, consist of read speech by professional newscasters, which has been claimed to be easier to annotate than more spontaneous speech (Mayo, Aylett, and Ladd 1997, 234), and they were annotated by trained and experienced transcribers only.

While newer studies report the kappa coefficient to quantify between-labeler consistency, the first systematic evaluation for ToBI assesses consistency in terms of percentage of matching transcriber-pair-words (Pitrelli, Beckman, and Hirschberg 1994). To calculate the percentage of matching transcriber-pair-words, they carried out pairwise comparisons for each word and each transcriber, accumulating the number of cases where any two transcribers agreed on a particular label (or nonlabel) for a word, and finally dividing this number by the total number of pairs where transcribers either agreed or disagreed. Using this measure, they reported 68.3 percent consistency for pitch accents, while Syrdal and McGory (2000) report 71 percent for their female corpus and 72 percent for their male corpus. Similar values of 71 percent were reported for GToBI pitch accents in German speech data (Grice et al. 1996) and slightly lower values for GlaToBI pitch accents in spontaneous Glaswegian English with 62 percent for nonexpert labelers and 69 percent for expert labelers (Mayo, Aylett, and Ladd 1997). Yoon et al. (2004) report a higher percentage of 86.57 percent, but this again refers to the consistency in labeling their reduced set of pitch accents. Insufficient transcriber reliability was the motivation for the development of ToBI Lite (Syrdal et al. 2000), reducing the set of pitch accent categories in American English to two (rising versus falling), which also served as the basis for the automatic recognition of these categories with high accuracy (the actual intertranscriber consistency was not reported).

Both kappa coefficients and transcriber/word pair accuracies suggest that intonation categories are more elusive than the categories in the segmental domain, where kappa values above 90 percent are not unusual (e.g., Gut and Bayerl 2004). In addition, to make use of these categories in intonation research, sufficiently large databases need to be available that are annotated accordingly. However, manual annotation of these categories is extremely time-consuming. Syrdal et al. (2001), for instance, found that experienced labelers take one hundred to two hundred times real time for annotating ToBI labels.

The fact that human labeling of intonation categories is both time-consuming and prone to labeler inconsistencies makes automatic labeling of these categories all the more attractive. One of the most interesting aspects of the PaIntE model is that it can actually be used to tackle this issue. Schweitzer and Möbius (2009) used the PaIntE parameters to predict pitch accents and boundary tones. They obtained accuracies of approximately 78 percent in the annotation of pitch accent types, and

accuracies of approximately 86 percent when addressing the annotation as a two-class problem, that is, when predicting presence versus absence of pitch accent rather than type of pitch accent. These results are slightly but probably not significantly better than those reported for read data by other recent studies (e.g., Hasegawa-Johnson et al. 2005; Sridhar, Bangalore, and Narayanan 2008; Rosenberg 2009).[8]

Unfortunately it is not valid to directly compare the accuracies reported herein to human labeler consistencies in terms of percentage of correct transcriber/word pairs, and studies on automatic labeling do not usually provide kappa values. However, to give an impression of the expected values, we used the Weka tool kit (Witten and Frank 2005) to calculate kappa values for the results reported in Schweitzer and Möbius (2009), obtaining $\kappa \approx 0.62$ for presence or absence and type of pitch accent. This indicates that there was a better consistency between the automatically predicted labels and the human gold standard labels than between human labelers in the study by Yoon et al. (2004), who reported $\kappa \approx 0.51$, but a lower consistency than that reported by Syrdal and McGory (2000) for experts' labels, which was $\kappa \approx 0.67$ and $\kappa \approx 0.69$ in a male and a female corpus, respectively.[9]

To conclude this section, we have argued here that the categories assumed by phonological models of intonation are more elusive than the categories in the segmental domain. Thus, two advantages of the PaIntE model are, first, that it does not depend on the assumption of such categories. Instead, the PaIntE parameters allow for quantifying established parameters such as peak height, peak alignment, or rise and fall amplitudes, on a continuous scale. This can also be exploited under an exemplar-theoretic account of intonation, as discussed in section 10.4.2. Second, if a phonological perspective is preferred, the PaIntE model can be used to automatically label intonation events with an accuracy close to that of human labelers.

10.6 Conclusion

The PaIntE model can be used, in an analysis mode, to approximate the shapes of natural F0 curves and, in a synthesis mode, to generate F0 contours that sound convincingly like natural ones. The model considers the intonation structure of an utterance as consisting of a sequence of intonation events, which can be mapped to elements of the linguistic structure, with simple contour interpolations between these events.

By default, PaIntE considers ToBI categories as relevant intonation events. This is an obvious choice, given the prevalence of the autosegmental model in intonation research. ToBI categories are, at least, a good approximation of salient intonational events. Mapping PaIntE parameter values to these categories therefore facilitates the comparison of results across otherwise different approaches in phonological and phonetic intonation research. Moreover, evidence of a correlation between ToBI-like categories and the PaIntE parameters was found by Möhler (2001), who reported that the acoustic distance between natural and generated F0 contours is smaller when GToBI(S) categories are taken into account than when parameter values based on generic clustering were used. However, if no annotation of intonation events is available, PaIntE parameters can be extracted for each syllable in a given utterance. This approach was taken, for instance, in first-language acquisition studies with young children whose productions cannot be described adequately by means of adult ToBI-like categories (Lintfert and Möbius 2012).

In this chapter, we have presented the motivation behind the PaIntE modeling approach and its mathematical formulation. The PaIntE function comprises a small

number of parameters with a linguistic interpretation whose values are estimated, learned, and generalized from speech databases. We explained the procedure of extracting the parameters from observed F0 contours and how to predict them, for instance, in the context of speech synthesis. Furthermore, we discussed the relation between the PaIntE intonation events and intonational categories posited by autosegmental approaches to intonation modeling. Finally, we presented several recent studies employing the PaIntE model. They show that PaIntE is a valuable contribution to intonation research beyond speech synthesis, which can serve to answer research questions both in the autosegmental tradition and in an exemplar-theoretic framework.

Notes

1. This holds under the assumption that establishing intonation categories works analogously to the segmental domain, where segmental categories are motivated by providing minimal pairs of words that differ only in the segmental category and have different meanings.

2. There is no closed-form expression for the true location of b, so the peak location must be approximated using numerical methods.

3. Also, the HH*L and H*M accents were not frequent enough in our data to reliably estimate their densities.

4. Thanks to Jörg Mayer for suggesting word finality as a possible explanation for the earlier peak.

5. https://librivox.org.

6. Note that research papers cited in this section are not always explicit about which version of the kappa coefficient (e.g., Cohen's kappa, Fleiss' kappa) they have used.

7. Perfect agreement is said to occur when $\kappa \geq 0.81$ (Landis and Koch 1977).

8. Comparing these results is straightforward because the data are similar: they all report results for the two-class problem, deriving the accent status from ToBI labels. All corpora consist of news-style read speech by professional speakers. However, it should be noted that the corpora are from two different languages with different ToBI systems.

9. We have only reported consistencies for pitch accents here; it should be noted that consistency for boundary tones is usually higher than for pitch accents, indicating that the boundary categories are easier to identify than pitch accents and in that respect are less problematic in our view.

References

Barbisch, M., G. Dogil, B. Möbius, B. Säuberlich, and A. Schweitzer. 2007. "Unit Selection Synthesis in the SmartWeb Project." In *Sixth ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, edited by P. Wagner, J. Abresch, S. Breuer, and W. Hess, 304–309. https://www.isca-speech.org/archive_open/ssw6/index.html.

Baumann, S., and M. Grice. 2006. "The Intonation of Accessibility." *Journal of Pragmatics* 38 (10):1636–1657.

Baumann, S., C. Röhr, and M. Grice, M. 2015. "Prosodische (de-)Kodierung des Informationsstatus." *Zeitschrift für Sprachwissenschaft* 34 (1):1–42.

Beckman, M. E. 1996. "The Parsing of Prosody." *Language and Cognitive Processes* 11 (1/2): 17–67.

Bruce, G. 1977. *Swedish Word Accent in a Sentence Perspective*. Lund, Sweden: Gleerup.

Büring, D. 1997. *The Meaning of Topic and Focus: The 59th Street Bridge Accent*. New York: Routledge.

Calhoun, S., J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver. 2010. "The NXT-Format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue." *Language Resources and Evaluation* 44 (4): 387–419.

Calhoun, S., and A. Schweitzer. 2012. "Can Intonation Contours Be Lexicalised? Implications for Discourse Meanings." In *Prosody and Meaning*, edited by G. Elordieta Alcibar and P. Prieto, 271–327. Berlin: Mouton DeGruyter.

Chen, A., E. den Os, and J. de Ruiter. 2007. "Pitch Accent Type Matters for Online Processing of Information Status: Evidence from Natural and Synthetic Speech." *Linguistic Review* 24:317–344.

Cooper, W. E., and J. M. Sorensen. 1981. *Fundamental Frequency in Sentence Production*. New York: Springer.

Dilley, L., and M. Brown. 2005. "The RaP Labeling System, v. 1.0." Unpublished manuscript. Michigan State University. http://speechlab.cas.msu.edu/RaP/RaP_Labeling_Guide_v1.0.pdf.

Dogil, G., and A. Schweitzer. 2011. "Quantal Effects in the Temporal Alignment of Prosodic Events." In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by W.-S. Lee and E. Zee, 595–598. City University of Hong Kong. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/.

Eckart, K., A. Riester, and K. Schweitzer. 2012. "A Discourse Information Radio News Database for Linguistic Analysis." In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 65–75. Berlin: Springer.

Féry, C. 1993. *The Meaning of German Intonational Patterns*. Tübingen, Germany: Max Niemeyer Verlag.

Féry, C., and F. Kügler. 2008. "Pitch Accent Scaling on Given, New and Focused Constituents in German." *Journal of Phonetics* 36 (4):680–703.

Godfrey, J., Holliman, E., and McDaniel, J. 1992. "Switchboard: Telephone Speech Corpus for Research and Development." In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520. https://doi.org/10.1109/ICASSP.1992.225858.

Goedemans, R. 1998. *Weightless Segments: A Phonetic and Phonological Study Concerning the Metrical Irrelevance of Syllable Onsets*. Leiden, the Netherlands: Holland Academic Grahics.

Goldinger, S. D. 1996. "Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:1166–1183.

Goldinger, S. D. 1997. "Words and Voices—Perception and Production in an Episodic Lexicon." In *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix, 33–66. San Diego: Academic Press.

Goldinger, S. D. 1998. "Echoes of Echoes? An Episodic Theory of Lexical Access." *Psychological Review* 105:251–279.

Goldsmith, J. A. 1976. "Autosegmental Phonology." PhD diss., MIT.

Grabe, E. 1998. "Pitch Accent Realizations in English and German." *Journal of Phonetics* 26:129–143.

Grice, M., and S. Baumann, S. 2002. "Deutsche Intonation und GToBI." *Linguistische Berichte* 191:267–298.

Grice, M., S. Baumann, and R. Benzmüller. 2005. "German Intonation in Autosegmental-Metrical Phonology." In *Prosodic Typology*, edited by S.-A. Jun, 53–83. Oxford: Oxford University Press.

Grice, M., M. Reyelt, R. Benzmüller, J. Mayer, and A. Batliner. 1996. "Consistency in Transcription and Labelling of German Intonation with GToBI." In *Proceedings of the Fourth International Conference on Spoken Language Processing*, edited by H. T. Bunnell and W. Idsardi, 1716–1719. https://www.isca-speech.org/archive/icslp_1996/.

Gussenhoven, C. 1984. *On the Grammar and Semantics of Sentence Accents*. Dordrecht, the Netherlands: Foris Publications.

Gussenhoven, C. 2005. "Transcription of Dutch Intonation." In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, 118–145. Oxford: Oxford University Press.

Gussenhoven, C., and A. Rietveld. 1988. "Fundamental Frequency Declination in Dutch: Testing Three Hypotheses" *Journal of Phonetics* 16:355–369.

Gut, U., and P. S. Bayerl. 2004. "Measuring the Reliability of Manual Annotations of Speech Corpora." In *Proceedings of Speech Prosody*, edited by B. Bel and I. Marlien, 565–568. https://www.isca-speech.org/archive/sp2004/.

Hasegawa-Johnson, M., K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, and T. Yoon. 2005. "Simultaneous Recognition of Words and Prosody in the Boston University Radio Speech Corpus." *Speech Communication* 46 (3–4):418–439.

House, D. 1996. "Differential Perception of Tonal Contours through the Syllable." In *Proceedings of the Fourth International Conference on Spoken Language Processing*, edited by H. T. Bunnell and W. Idsardi, 2048–2051. https://www.isca-speech.org/archive/icslp_1996/.

House, D. 1997. "Temporal Alignment Categories of Accent-Lending Rises and Falls." In *Proceedings of Eurospeech*, edited by G. Kokkinakis, N. Fakotakis, and E. Dermatas, 879–882. https://www.isca-speech.org/archive/eurospeech_1997/.

Jilka, M., G. Möhler, and G. Dogil. 1999. "Rules for the Generation of ToBI-Based American English Intonation." *Speech Communication* 28:83–108.

Johnson, K. 1997. "Speech Perception without Speaker Normalization: An Exemplar Model." In *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix, 145–165. San Diego: Academic Press.

Kelly, N., and K. Schweitzer. 2015. "Examining Lexical Tonal Contrast in Norwegian Using Intonation Modelling." In *Proceedings of the Eighteenth International Congress of Phonetic Sciences*. Glasgow: University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/proceedings.html.

Kelly, N. E., and R. Smiljanić. 2014. "The Effect of Focus on Norwegian Tonal Accent." In *Proceedings of the Fourth International Symposium on Tonal Aspects of Languages*, edited by C. Gussenhoven, Y. Chen, and D. Dediu, 95–99. https://www.isca-speech.org/archive/tal_2014/.

Kohler, K. J. 1990. "Macro and Micro F0 in the Synthesis of Intonation." In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. E. Beckman, 115–138. Cambridge: Cambridge University Press.

Kohler, K. 1991. "A Model of German Intonation." In *Studies in German Intonation, Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel* 25:295–360.

Kollmeier, B., T. Brand, and B. T. Meyer. 2008. *Perception of Speech and Sound*. Berlin: Springer.

Kristoffersen, G. 2006. "Tonal Melodies and Tonal Alignment in East Norwegian." In *Nordic Prosody IX*, edited by G. Bruce and M. Horne, 157–166, Frankfurt: Peter Lang.

Lacerda, F. 1995. "The Perceptual-Magnet Effect: An Emergent Consequence of Exemplar-Based Phonetic Memory." *Proceedings of the Thirteenth International Congress of Phonetic Sciences,* edited by K. Elenius and P. Branderud. 2:140–147. Stockholm: KTH and Stockholm University.

Ladd, D. R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R., I. Mennen, and A. Schepman. 2000. "Phonological Conditioning of Peak Alignment in Rising Pitch Accents in Dutch." *Journal of the Acoustical Society of America* 107 (5): 2685–2696.

Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1):159–174.

Lintfert, B., and B. Möbius. 2012. "Describing the Development of Intonational Categories Using a Target-Oriented Parametric Approach." In *Proceedings of Interspeech 2012*, 1965–1968. https://www.isca-speech.org/archive/interspeech_2012/.

Mayer, J. 1995. *Transcription of German Intonation—the Stuttgart System*. Technical report. Stuttgart, Germany: Institute of Natural Language Processing, University of Stuttgart.

Mayo, C., M. Aylett, and D. R. Ladd. 1997. "Prosodic Transcription of Glasgow English: An Evaluation Study of GlaToBI." In *Proceedings of an ESCA Workshop: Intonation: Theory, Models and Applications*, edited by G. Kouroupetroglou and G. Carayiannis, 231–234.

Möhler, G. 2001. "Improvements of the PaIntE Model for F0 Parametrization." Unpublished manuscript. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.6756.

Möhler, G., and A. Conkie. 1998. "Parametric Modeling of Intonation Using Vector Quantization." In *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, 311–316. https://isca-speech.org/archive_open/ssw3/.

Ostendorf, M., P. Price, and S. Shattuck-Hufnagel. 1996. *Boston University Radio Speech Corpus*. Philadelphia: Linguistic Data Consortium.

Peters, J. 2014. *Intonation*. Heidelberg: Winter.

Pierrehumbert, J. 1979. "The Perception of Fundamental Frequency Declination." *Journal of the Acoustical Society of America* 66:363–369.

Pierrehumbert, J. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pierrehumbert, J. 1981. "Synthesizing Intonation." *Journal of the Acoustical Society of America* 70:985–995.

Pierrehumbert, J. 2001. "Exemplar Dynamics: Word Frequency, Lenition and Contrast." In *Frequency and the Emergence of Linguistic Structure*, edited by J. Bybee and P. Hopper, 137–157. Amsterdam: Benjamins.

Pierrehumbert, J. 2003. "Probabilistic Phonology: Discrimination and Robustness." In *Probability Theory in Linguistics*, edited by R. Bod, J. Hay, and S. Jannedy, 177–228. Cambridge, MA: MIT Press.

Pierrehumbert, J., and J. Hirschberg. 1990. "The Meaning of Intonational Contours in the Interpretation of Discourse." In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Pitrelli, J., M. Beckman, and J. Hirschberg. 1994. "Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework." In *Proceedings of the Third International Conference on Spoken Language Processing*, 123–126. https://www.isca-speech.org/archive/icslp_1994/.

Rapp, S. 1998. "Automatisierte Erstellung von Korpora für die Prosodieforschung." PhD diss., University of Stuttgart.

Rosenberg, A. 2009. "Automatic Detection and Classification of Prosodic Events." PhD diss., Columbia University.

Schweitzer, A. 2011. "Production and Perception of Prosodic Events—Evidence from Corpus-based Experiments." PhD diss., Universität Stuttgart.

Schweitzer, A., and B. Möbius. 2009. "Experiments on Automatic Prosodic Labeling." In *Proceedings of the Tenth Annual Conference of the International Speech Communication Association*, 2515–2518. https://www.isca-speech.org/archive/interspeech_2009/.

Schweitzer, K., M. Walsh, S. Calhoun, H. Schütze, B. Möbius, A. Schweitzer, and G. Dogil. 2015. "Exploring the Relationship between Intonation and the Lexicon: Evidence for Lexicalised Storage of Intonation." *Speech Communication* 66:65–81.

Sproat, R., ed. 1998. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Dordrecht, the Netherlands: Kluwer.

Sridhar, V. K. R., S. Bangalore, and S. Narayanan. 2008. "Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework." *IEEE Transactions on Audio, Speech, and Language Processing* 16 (4): 797–811. https://doi.org/10.1109/TASL.2008.917071.

Syrdal, A. K., J. Hirschberg, J. McGory, and M. Beckman. 2001. "Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody." *Speech Communication* 33:135–151.

Syrdal, A. K., and J. McGory. 2000. "Inter-Transcriber Reliability of ToBI Prosodic Labeling." *Proceedings of the Sixth International Conference on Spoken Language Processing* 3:235–238. https://www.isca-speech.org/archive/icslp_2000/.

Syrdal, A. K., C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K.-S. Lee, and M. J. Makashay. 2000. "Corpus-Based Techniques in the AT&T NextGen Synthesis System." *Proceedings of the Sixth International Conference on Spoken Language Processing* 3:410–413. https://www.isca-speech.org/archive/icslp_2000/.

Taylor, P. 1998. "The Tilt Intonation Model." In *Proceedings of the Fifth International Conference on Spoken Language Processing*, 1383–1386. https://www.isca-speech.org/archive/icslp_1998/.

Taylor, P. 2009. *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

Taylor, P., R. Caley, A. W. Black, and S. King. 1999. *Edinburgh Speech Tools Library*. System Documentation Edition 1.2, for 1.2.0. June 15. http://festvox.org/docs/speech_tools-1.2.0/.

Terken, J., and J. Hirschberg. 1994. "Deaccentuation of Words Representing 'Given' Information: Effects of Persistence of Grammatical Role and Surface Position." *Language and Speech* 37:125–145.

van Santen, J. P. H., and B. Möbius. 2000. "A Quantitative Model of F0 Generation and Alignment." In *Intonation: Analysis, Modelling and Technology*, edited by A. Botinis, 269–288. Dordrecht, the Netherlands: Kluwer.

Wade, T., G. Dogil, H. Schütze, M. Walsh, and B. Möbius. 2010. "Syllable Frequency Effects in a Context-Sensitive Segment Production Model." *Journal of Phonetics* 38 (2): 227–239.

Wahlster, W. 2004. "SmartWeb: Mobile Applications of the Semantic Web." In *KI 2004: Advances in Artificial Intelligence*, edited by S. Biundo, T. Frühwirth, and G. Palm, 50–51. Berlin: Springer.

Walsh, M., B. Möbius, T. Wade, and H. Schütze. 2010. "Multilevel Exemplar Theory." *Cognitive Science* 34:537–582.

Witten, I. H., and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann.

Yoon, T.-J., S. Chavarría, J. Cole, and M. Hasegawa-Johnson. 2004. "Intertranscriber Reliability of Prosodic Labeling on Telephone Conversation Using ToBI." In *Proceedings of the Eighth International Conference on Spoken Language Processing*, 2729–2732. https://www.isca-speech.org/archive/interspeech_2004/.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data is available.
Names: Barnes, Jonathan, 1970– editor. | Shattuck-Hufnagel, Stefanie, editor.
Title: Prosodic theory and practice / edited by Jonathan Barnes and Stefanie Shattuck-Hufnagel.
Description: Cambridge, Massachusetts : The MIT Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021000764 | ISBN 9780262543170 (paperback)
Subjects: LCSH: Prosodic analysis (Linguistics)
Classification: LCC P224 .P739 2022 | DDC 414/.6—dc23
LC record available at https://lccn.loc.gov/2021000764