Prosodic Theory and Practice

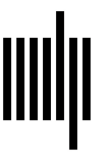
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel

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

7 The Prosogram Model for Pitch Stylization and Its Applications in Intonation Transcription

Piet Mertens

7.1 Introduction

This chapter describes a generic, integrated approach to the analysis of speech prosody, covering its acoustic, perceptual, and linguistic manifestations, as well as the relationships and mapping among them. Analysis proceeds bottom-up, starting from acoustic parameters, over pitch stylization simulating tonal perception, via the labeling of pitch levels and movements using discrete symbols, so as to ultimately obtain a structured representation of intonation in which pitch events are aligned with prosodic structure, forming prosodic units, with their internally structured pitch contours. Both the initial analysis and its more advanced applications are implemented as a computational system, allowing for a systematic validation of the model and its components.

Prosodic features of speech may be characterized at three major levels of observation: the acoustic, the perceptual, and the linguistic. The last is commonly referred to as the phonological level, because it involves abstraction and reduction of phonetic detail. For completeness, a fourth characterization of prosody should be added, which is the physiological level. Whatever we measure, perceive, or categorize is first produced by the vocal apparatus—by neural instructions to the muscles in the thorax and the larynx. These physiological aspects, however, are not dealt with in this chapter.

Let us briefly sketch the three representations discussed here. First, *acoustic analysis* measures continuous physical parameters of the speech signal, such as fundamental frequency (F0), voicing, intensity, and spectral energy distribution. Spectral information changing over time enables the identification of sounds, syllables, and pauses. Second, perceptual processing of these acoustic events results in an *auditory representation*, which entails a substantial reduction of the variation present in the acoustic signal and produces a segmentation into syllable-sized units. Due to this segmentation, the continuous speech signal is transformed into a sequence of short duration fragments. This process has a crucial impact on the perception of prosodic attributes such as pitch, loudness, and duration. For instance, F0 change is transformed into a sequence of pitch steps and pitch movements. Also, the emergence of syllables gives rise to derived prosodic attributes such as prominence, speech rate, and rhythm. The third observation level, the *phonological characterization*, assumes a minimal set of distinctive prosodic forms (called "tones" in some approaches), associated with either individual syllables or sequences of syllables. The concatenation of such forms results in contours anchored at stressed syllables. These tones and contours are assumed to have a function in speech communication. Although related, the three representations or observation levels differ considerably: one abstract phonological form may correspond at the perceptual level to several forms with slightly different auditory shapes. In

addition, the acoustic realization of such forms depends on the phonetic composition of the syllable on which it occurs, it varies with speech rate, and so forth.

When comparing these three representations of prosody for random utterances, the differences among them are striking. Moving bottom-up from acoustics to phonology, one goes from continuous parameters over sequences of simplified shapes to abstract forms noted by discrete symbols. Although different, each of these representations is adequate in some respect and appropriate for some purpose. Acoustic specifications constitute the objective representation of the sound signal, from which all other representations are derived. They are essential in speech synthesis, for instance. But to understand speech, listeners do not need spectrograms or pitch tracks; auditory information suffices. The perceptual representation, as witnessed by handmade transcriptions of prosody, based on auditory information only, shows to what extent perceptual processing transforms the acoustic information, smoothing pitch variation and introducing segmentation, thereby creating a pattern in auditory memory that the listener can describe verbally and reproduce in a graphical form. Finally, the abstract linguistic representation further categorizes these auditory shapes, eliminating free and contextual variation, normalizing pitch range differences between speakers, normalizing temporal aspects such as speech rate and rhythm, generalizing over syllable sequences of various lengths, and introducing structure based on word stress, for example.

A comprehensive theory of speech prosody should aim not only at an adequate acoustic representation of the prosodic parameters or a compact (or even a minimal) characterization of distinctive forms. It should also aim at explaining the mapping between representations of increasing degrees of abstraction. How does the perceived auditory pattern emerge from the acoustic data? Which mechanisms account for the categorization of the auditory shapes into a closed set of basic forms, which we transcribe by discrete symbols? Such conversions are essential, because understanding and simulating these mappings show they are reproducible and predictable. Hence, the (continuous) perceptual representation and the (discrete) linguistic representation are not arbitrary but, rather, have a cognitive status.

The general approach outlined is illustrated in figure 7.1. Its three panels broadly match the acoustic, perceptual, and linguistic observation levels discussed. The top panel shows the acoustic level, illustrated by the F0 trace. The middle panel, corresponding to the perceptual level, shows the stylized pitch curve aligned with the syllable chain, together with other parameters used in the stylization (e.g., intensity, voicing, phoneme and syllable boundaries, and pauses). Finally, the bottom panel shows a discrete transcription using symbols for pitch levels (low, mid, high, bottom, and top) and pitch movements (rise, fall, level). For instance, the symbol HF, for "high fall," indicates that the corresponding syllable contains a large downward pitch movement starting from a high pitch level.

Every symbolic transcription of intonation makes theoretical assumptions about its underlying structure: its basic elements and their possible combinations, their function in language, and their interaction with the units and syntactic organization of the segmental layer. The impact of such theoretical assumptions is demonstrated by the bewildering number of competing intonation models that have been proposed. The distinctive feature of the approach presented here is that it proposes a comprehensive prosody model, combining the acoustic, perceptual, and discrete symbolic specifications, as well as the mapping between successive representation levels. Most importantly, the perceptual and symbolic representations of figure 7.1 were computed automatically from the speech signal and the phonetic alignment, strongly supporting the claim that

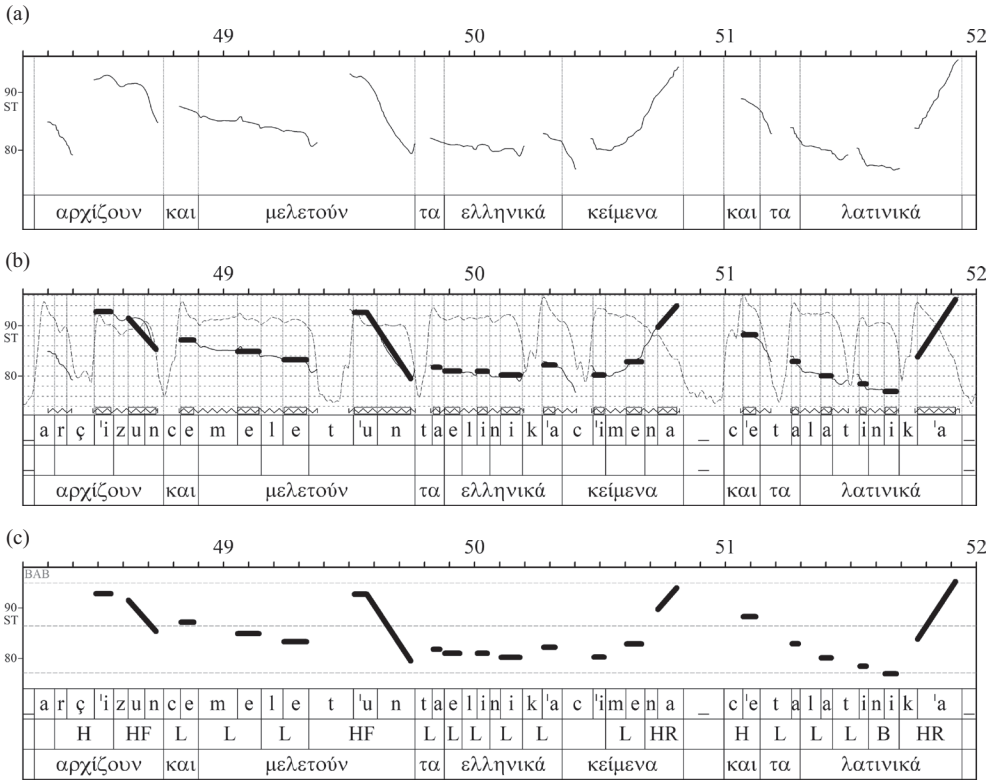


Figure 7.1
Acoustic, perceptual, and symbolic representations of the pitch contour of the Greek utterance "αρχίζουν και μελετούν τα ελληνικά κείμενα και τα λατινικά" (They start studying the Greek texts and the Latin ones), pronounced by a male speaker. The upper panel shows fundamental frequency on a semitone scale. The bold line in the center panel shows the stylized pitch based on a model simulating tonal perception. The central panel also includes acoustic parameters used by the stylization, such as intensity (dashed line) and voicing (sawtooth line), as well as the segmentation into sonority peaks (shown as boxes). The three annotation layers (or "tiers") in the central panel indicate, from top to bottom, phonetic symbols (including stress marks), syllables, and words. The bottom panel shows a symbolic transcription of pitch levels and movements in the second tier (aligned with the syllables). The detected pitch range is visualized by the three horizontal dashed lines, indicating top and bottom of the pitch span and the median pitch.

by simulating auditory perception and, in particular, tonal perception, the acoustic signal may be mapped onto its perceptual representation, which in turn may be categorized into discrete symbols. The latter are compatible with those posited in linguistic models lacking a computational model. The mapping in the opposite direction, from discrete symbols to the acoustic speech signal, has already been illustrated in text-to-speech synthesis (Mertens et al. 2001).

The remainder of this chapter is organized as follows. Section 7.2 deals with pitch contour stylization simulating pitch perception. After describing the pitch perception in speech (section 7.2.1), it outlines the Prosogram algorithm for simulating it (section 7.2.2) and discusses the validation of this model (section 7.2.3). Section 7.3

summarizes the prosodic attributes that are computed within this computer program for each syllable, as well as for sequences of syllables. One such attribute, pitch range, is discussed in more detail. These data may be used in experimental research and speech applications. The next two sections present two applications of the Prosogram model in the area of prosody transcription: in section 7.4, the automatic transcription of pitch levels and movements observed at the level of the syllable, and in section 7.5, fully fledged intonation transcription, making explicit underlying structure, prosodic units, and pitch contours. The final section (7.6) discusses the major assumptions underpinning this approach.

Before we set out, a quick reminder on pitch scales. In voiced speech, the *fundamental frequency, F0*, of the quasi-periodic waveform corresponds to the frequency of vocal fold vibration. It is an acoustic property. *Pitch* is the perceptual correlate of F0. Whereas frequency is measured in hertz, various scales are used for pitch measurement (Nolan 2003): hertz, semitone (ST), Mel, Bark, equivalent regular bandwidth (Hermes and Van Gestel 1991), and octave-median (De Looze and Hirst 2014). The musical scale in ST indicates a *pitch interval*, that is, the distance in pitch between two points in time (for instance, between two musical notes, two sounds, or the start and end of a pitch glide), rather than an absolute pitch or frequency. The distance D in ST between two frequencies f_1 and f_2 is given by: $D = 12 \log_2 (f_2/f_1)$ (Baken and Orlikoff 2000, 148). When f_2 is twice as high as f_1, they are separated by one octave. The ST scale divides the octave into twelve equal steps on a logarithmic scale. When speakers with a different vocal range (for instance, a man and a woman) pronounce the same utterance while imitating its intonation, their fundamental frequency range will probably differ, although the pitch intervals (in ST) used in their intonation contours may be similar. The ST scale thus allows for pitch contour normalization, maintaining the actual pitch intervals independently of absolute pitch. The ST scale may also be used as an absolute scale, by setting f_1 in the formula for D to a fixed reference frequency. (The illustrations in this chapter use ST relative to 1 Hz. The reference frequency may be selected arbitrarily: a given pitch interval remains the same size in ST, whether the reference frequency is set to 1 Hz or 100 Hz. For more details on the ST scale, see Mertens (2020, section 9.1).

7.2 Pitch Stylization Based on a Tonal Perception Model

Stylization is a process that simplifies the F0 curve of the speech signal. It removes local F0 variations which are perceptually irrelevant, because they are too small or too short to be perceived, because they occur in spectrally unstable regions, such as the transition between consonants and vowels (micro-prosody or micro-intonation), and so forth. By doing so, it maintains only those aspects that potentially have a function in speech communication. The type of stylization described here aims at simulating the auditory representation as it emerges from perceptual processing. In this way, it helps to reveal the underlying intonation pattern.

In general, pitch stylization makes the assumption that pitch contours of utterances can be adequately synthesized and hence be represented by a sequence of basic shapes, such as straight lines or curves. In his detailed account of stylization, Hermes (2006) distinguishes two major approaches, depending on the nature of the elementary parts. The first approach obtains a sequence of elementary shapes, either straight lines ('t Hart, Collier, and Cohen 1990; Scheffers 1988; Spaai et al. 1993) or curves (Taylor 1994, 2000; Hirst, Nicolas, and Espesser 1991; Hirst et al. 2000). In this approach, the turning points of the F0 contour determine the boundaries between elementary shapes.

In the second approach, the stylization takes the form of a sequence of syllabic tones, that is, simple or compound pitch movements (including steady pitch) associated with syllables (d'Alessandro and Mertens 1995; Mertens 2004a, 2004b). Obviously, syllabic tones are also elementary shapes. But whereas in the first approach, the pitch contour is seen as a continuum and the boundaries of elementary units are determined by pitch change only, the second approach acknowledges the impact of spectral change on tonal perception. As a result, the segmentation based on pitch change is preceded by a segmentation based on spectral change.

A closer look at the first major approach shows that some of these models implicitly take into account spectral change and amplitude variation. This is illustrated by the IPO approach ('t Hart, Collier, and Cohen 1990) (IPO refers to the research institute "Instituut voor Perceptie Onderzoek" in Eindhoven, the Netherlands, where this approach was developed), which distinguishes early and late pitch movements depending on the temporal alignment of the movement with vowel onset and vowel end (Hermes 1987). But other models, such as the quadratic spline modeling (Hirst, Nicolas, and Espesser 1991), claim that the perception of pitch variation in speech is continuous (Hirst 2011) and reject any form of segmentation. In such a view, turning points may occur anywhere in the signal, within vowels or consonants, or even in silent portions.

The remainder of this section takes a closer look at pitch stylization based on tonal perception. It lists various perceptual phenomena observed in psychoacoustics, which are simulated in the stylization model, as well as in the Prosogram tool, which is an implementation of this model using the Praat speech analysis program (Boersma and Weenink 2012).

7.2.1 Tonal Perception

7.2.1.1 Spectral instability and its impact on pitch perception Spectral change and amplitude drop at the transition between sounds (consonants and vowels) can perceptually mask F0 change during this transition ('t Hart, Collier, and Cohen 1990, 36). Experiments by House (1990) show that an F0 variation—for instance, a linear rise of fixed duration and size—is perceived differently, as either a pitch movement or an abrupt change of pitch, depending on its location relative to the vowel onset in the syllable. This phenomenon (known as the *spectral stability hypothesis*) is attributed to spectral change at the transition between sounds, which lowers sensitivity to pitch change. The amount of spectral change at sound transitions depends on the nature of the sounds involved; nasals and liquids are more similar to vowels than fricatives, for instance. A more detailed segmentation model based on spectral change also takes into account the nature of the sounds in the syllable rhyme (House 1995, 1996). This perceptual phenomenon explains why hearers are able to distinguish pitch movements occurring between adjacent syllables from those occurring within one and the same syllable.

7.2.1.2 Intrasyllabic pitch movement and the glissando threshold Some syllables are perceived with a stable pitch, others with a changing pitch. In music, the latter type would be called a *glissando*. For a gliding pitch, the observed F0 change will exceed a threshold, which depends on the pitch interval covered by the glide and its duration ('t Hart, Collier, and Cohen 1990). Listening experiments with natural or synthetic speech (Rossi, 1971; Rossi et al. 1981; Mertens, Beaugendre, and d'Alessandro 1997) show that many intrasyllabic F0 variations observed in speech are indeed perceived

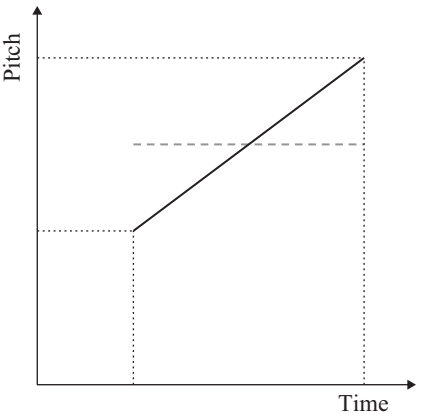


Figure 7.2
Effect of the GT for an F0 variation (solid line) with a given duration and size (dotted lines). When the combination of duration and size is lower than the GT, the F0 variation is perceived as a sound with stable pitch (dashed line).

as glissandi. To decide whether a syllable is perceived with a stable or a gliding pitch, the combination of duration and size of the corresponding F0 variation is compared to the glissando threshold (GT), as illustrated for a uniform F0 change in figure 7.2. When they exceed the GT, a gliding pitch is obtained (solid line), otherwise a stable pitch is perceived with a frequency approaching the median F0 of the variation (dashed line).

The GT was first observed and measured in experimental settings, using stimuli with known properties, such as pure tones or synthetic vowels, presented repeatedly, separated by silence, hence followed by a pause. Natural, continuous speech is characterized instead by spectral instability due to sound concatenation and by the occasional presence of pauses. As a result, the threshold is higher in connected speech than for isolated sounds. House (1995) shows that the presence of the pause facilitates pitch perception—in other words, it lowers the GT. (For a detailed description of the measurement of the GT, see d'Alessandro and Mertens 1995, section 2.5.2.)

7.2.1.3 Simple and compound pitch shapes Hearers have the ability to discriminate between simple and compound pitch movements occurring within a syllable. For compound movements, the successive parts, or tonal segments, may have slopes of opposite direction (e.g., a rise followed by a fall) or of same direction but different rate (e.g., slow rise followed by fast rise, or a level part followed by rise). Whether a pitch variation is perceived as two or more successive pitch movements rather than as a single movement depends on the amount of slope change between the successive tonal segments (see 't Hart, Collier, and Cohen 1990). In figure 7.3 the solid line shows a compound pitch variation consisting of two uniform parts of different slope. When the slope change is sufficiently large, the pitch variation is perceived as a sequence of two parts of different slope, otherwise as one part with a uniform slope, corresponding to the dashed line. In d'Alessandro and Mertens (1995) and in the Prosogram tool, this phenomenon is modeled by the differential glissando threshold (DGT), expressed as $g_2 - g_1$, with g_1 and g_2 the slopes in ST/s of the first and the second tonal segment. (For the DGT measure, see d'Alessandro and Mertens 1995, section 2.5.3.)

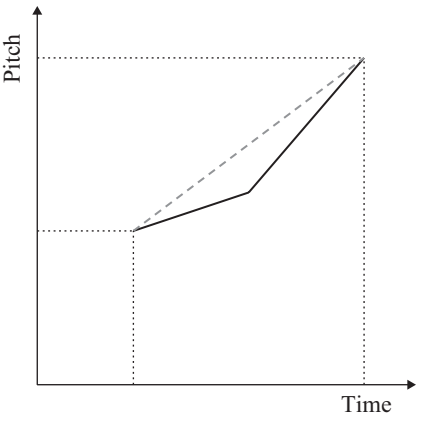


Figure 7.3
Effect of the DGT for a compound F0 variation (solid line) with a given duration and size (dotted lines). When the slope change exceeds the DGT, the variation is perceived as a sequence of two parts with a different slope, otherwise as a single rising pitch movement with constant slope (dashed line).

7.2.1.4 Minimal duration of tonal segments In addition to the two thresholds just mentioned, there appears to be an additional constraint on the minimal duration of a tonal segment. For compound pitch movements, consisting of two or more parts of uniform slope and with sufficiently large slope changes between them (i.e., exceeding the DGT), each segment needs to have a minimal duration of about 0.035 s. A shorter segment is not perceived as an individual tonal segment but is perceptually merged with the adjacent segment that is most similar.

7.2.1.5 Temporal alignment of pitch gestures with speech sounds Pitch events that are perceptually similar vary considerably in their acoustic shape, depending on the nature of the sounds over which they occur. We are not referring here to micro-prosody (also called micro-intonation), which among other things includes F0 variation due to sound nature (e.g., vowel aperture or consonant voicing, so-called *intrinsic micro-prosody*) or phonetic context (e.g., the pitch perturbation during vowel onset after voiceless obstruents, so-called *co-intrinsic micro-prosody*). Rather, we refer to the way in which a pitch event (a target or a movement) is temporally aligned with the segmental layer. Consider a high pitch target on a vowel, following a low pitch target on the preceding syllable. The transition from low to high may show up as either a pitch step (a discontinuity) or a pitch rise during syllable onset, depending on the voicing of the consonants in the onset. When the syllable onset contains highly sonorous sounds, such as a nasal or a glide, this usually results in a slower rise, spread over the onset and the vowel. So, a given pitch target sequence results in quite different F0 contours in the syllable onset, depending on phonetic context. Except for trained phoneticians, listeners generally are not aware of such acoustic differences. Obviously, similar observations can be made for the syllable rhyme, where the pitch movement will be spread over the sonorous segments in the coda when present.

Although the impact of intensity and spectral change on the perception of pitch events was acknowledged in earlier research (e.g., Rossi 1978), such phenomena have largely been ignored in dominant analysis frameworks since the 1980s. The IPO approach

('t Hart, Collier, and Cohen 1990) reduces intonation to a concatenation of standardized linear pitch movements (in the log F0 versus time domain) anchored at a syllable (either early, late, or very late pitch movements) or a prosodic boundary. The variation due to phonetic context was considered perceptually irrelevant. Conversely, in the autosegmental framework (for an overview, see Grice 2006 or Ladd 2008), intonation is characterized as a sequence of tones (or pitch levels) associated with particular syllables (called "tone-bearing units"), and the pitch of other speech portions is supposed to be accounted for by interpolation between these targets. Recently, however, segmental context has gained renewed attention in the context of the tonal center of gravity model (Barnes et al. 2014) of the perceptual integration of pitch change, weighted by segment sonority.

Given the acoustic variability of perceptually equivalent pitch contours, a prosody model should explicitly take into account the temporal alignment of pitch events with the segmental layer.

7.2.2 The Stylization Algorithm

This section describes how the above-enumerated perceptual phenomena are simulated in the stylization algorithm.

7.2.2.1 Segmentation into syllabic nuclei or rhymes The *syllable* is a central unit for the characterization of prosodic attributes such as prominence, stress, syllable length, speech rate, rhythm, and pitch movements. While some pitch events are synchronized with individual syllables (in particular, the stressed syllable), others are carried by a sequence of contiguous unstressed syllables ('t Hart 1998). Even for such gradual movements, the boundaries of the movement coincide with those of a syllable. Hence, the analysis of prosody requires identifying syllable boundaries.

The syllable, however, is a phonological unit. From an acoustic perspective, the exact location of syllable boundaries is often unclear. Whereas plosives and voiceless fricatives are characterized by abrupt changes in the acoustic signal, boundaries are acoustically less clear for glides, nasals, laterals, approximants, and diphthongs. Also, from a production perspective, sounds may be ambisyllabic: in this case, the closure of a consonant is part of a syllable's coda, whereas its release starts the next syllable's onset. For pitch perception, what matters is not the boundaries of the phonological syllable but rather those of the central part of a syllable, characterized by high sonority and relative spectral stability. This central part will be referred to as the *syllabic nucleus*. (Note that this term is already used in phonology to designate the vowel part of a syllable.)

In addition, fundamental frequency is defined only for voiced sounds, and intensity and spectral distribution may vary considerably during the syllable. These changes in periodicity, intensity, and spectrum affect the perception of pitch changes (Rossi 1978).

The procedure for syllabic nucleus detection adopted by Prosogram simulates perceptual properties while using acoustic parameters. The unit is identified as the part of the voiced portion of a syllable, located around the intensity peak within the vowel, where intensity decrease is below a given threshold. For the left boundary, this threshold is at 2 dB below the peak. For most recording conditions, this provides an acceptable approximation of the vowel onset. For the right boundary, the threshold is chosen so as to include the voiced, high-intensity part of the rhyme. More precisely, the intensity drop threshold is equal to 80 percent of intensity difference between the peak and the local minimum in the time interval between the peak and syllable end, with a minimum of 3 dB. When the resulting time interval contains an abrupt pitch change, most often resulting from an octave jump pitch detection error, the interval is

truncated at this point to avoid erroneous pitch values. A minimal duration (0.025 s) for the resulting syllabic nucleus is also required.

For best accuracy, the segmentation may be guided by the phonetic and syllabic alignment provided by the corpus annotation. In this case, these alignments are used to locate the time window where the syllabic nucleus is to be found, more precisely to locate the vowel and the syllable rhyme, consisting of the vowel and the coda. The resulting segmentation is illustrated in figure 7.4.

In addition, the Prosogram tool provides alternative segmentation types, which may be selected depending on the temporal alignments available for the corpus: phonemes only, syllables only, or vowels only.

For corpora without phonetic alignment, automatic segmentation is available, which is based on intensity of band-pass filtered signal (300–3,500 hertz), intensity, voicing, and F0 discontinuities. This is illustrated in figure 7.5.

7.2.2.2 Pause detection Silent pauses play a role in discourse—for instance, by marking the boundaries between major syntactic constituents or in speaker turn management. Silent pauses also affect the perception of pitch events by lowering the GT (House 1995).

Pause detection may be complicated by background noise and speech dynamics, which depend on recording conditions and vary from one corpus to the next. Most pause detection algorithms exploit the difference in intensity between voiced speech and silent pauses. In case of background noise, the intensity of nonspeech increases. In case of changing recording conditions, such as gain adjustments during recording, including automatic gain control, it is difficult to select an optimal threshold.

As a practical solution, the following procedure is used. The segmentation into syllables results in a sequence of syllabic nuclei. The gap between the end of one nucleus and the beginning of the next provides a rough estimate of pause length. When this gap exceeds 0.35 s, it is interpreted as a pause. In the illustrations, detected pauses are marked by a *P* at the start time of the gap.

7.2.2.3 Pitch stylization Stylization is applied to the F0 data of each syllabic nucleus obtained by the segmentation (see section 7.2.2.1). The algorithm, which is described in detail in d'Alessandro and Mertens (1995), may be summarized as follows. For each syllabic nucleus, the pitch contour is divided into one or more parts of uniform slope (either level, rising or falling) called *tonal segments*. First, turning points in the pitch contour are located by order of importance. For a given time interval, a candidate turning point is found at the time of maximum distance between the observed F0 and a straight line connecting the F0 values at the start and end of the interval under consideration. This candidate turning point is accepted only when all following conditions are met: (i) the F0 distance at the turning point exceeds 1 ST, (ii) both resulting segments exceed the minimal duration of a tonal segment, (iii) at least one of the segments is a glissando, and (iv) the slope change at the turning point exceeds the DGT. The same procedure is applied recursively to the time intervals on both sides of a valid turning point. Recursion halts when no additional turning point is found. After the slicing into tonal segments, the perceived pitch at the turning points is estimated. If the pitch variation of a tonal segment is below the GT, it is replaced by a level slope, with a pitch equal to the median F0 in the corresponding time interval.

In the current model, perceptual thresholds are dynamically adjusted for continuous speech. For a syllable preceding a pause, the GT is set to the value for isolated

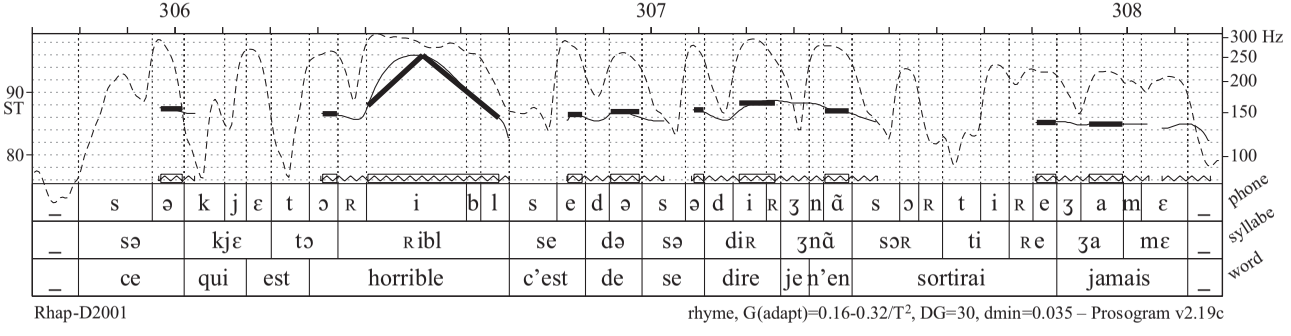


Figure 7.4

Illustration of segmentation into syllabic nuclei guided by phonetic alignment, for the French utterance, "Ce qui est horrible c'est de se dire je n'en sortirai jamais" (What is horrible is to say to oneself: I will never get out of this situation), taken from the Rhapsodie corpus (recording Rhap-D2001; Lacheret-Dujour et al. 2019). The upper part shows the acoustic parameters of intensity (the dashed line), voicing (sawtooth), and fundamental frequency (thin line), mostly covered by the thick line of the stylized pitch. Pitch is plotted on an ST scale (relative to 1 Hz), with horizontal calibration lines at two ST steps. The lower part shows various corpus annotation tiers: phonetic alignment, syllables, and words. The syllabic nuclei appear as small boxes on top of the sawtooth voicing line. For the syllable [ʀibl], the nucleus includes the vowel and part of the coda. The syllables [kjɛ], [sɔʀ], and [ti] are analyzed as unvoiced; consequently no nucleus is detected.

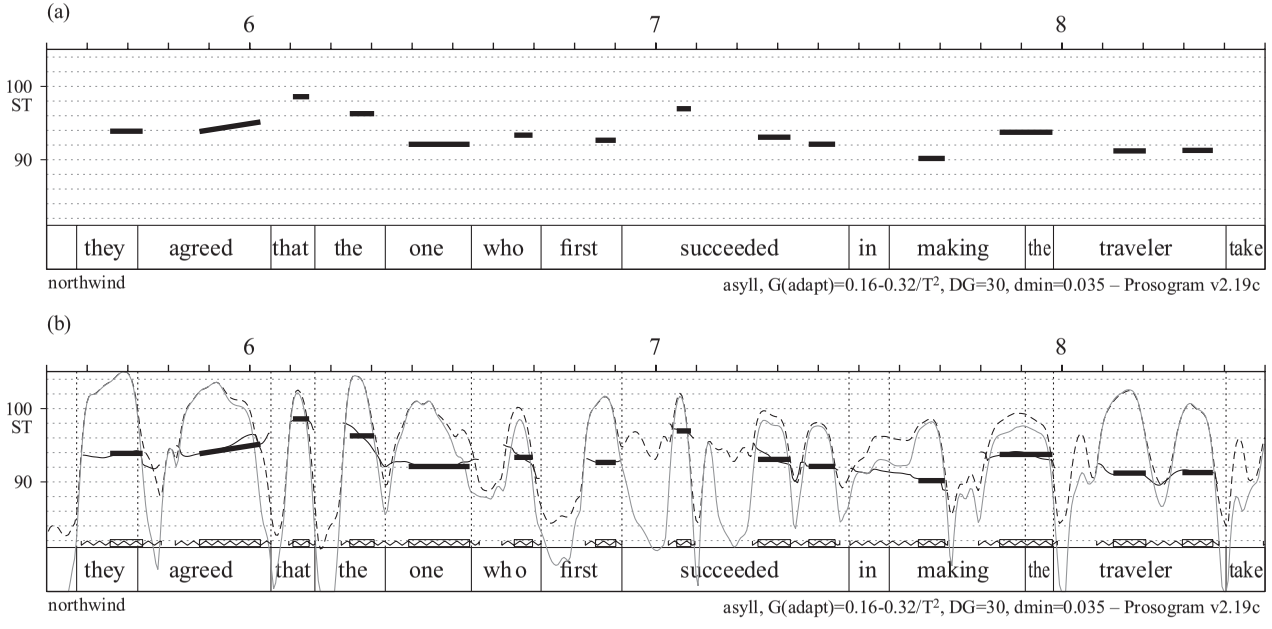


Figure 7.5
Illustration of the automatic segmentation into syllabic nuclei based on acoustic information only, for the utterance, "They agreed that the one who first succeeded in making the traveler take . . . ," from "The North Wind and the Sun" passage in American English (International Phonetics Association). The upper panel shows the stylized pitch of syllabic nuclei and the word annotation (not used in the segmentation process). The lower panel also shows the acoustic parameters of intensity (dashed line), intensity of the band-pass filtered signal (continuous solid line with larger dips), voicing (sawtooth), and F0 (thin line interrupted at unvoiced portions, mostly covered by the bold line of the stylization). Pitch is plotted on an ST scale, with horizontal calibration lines at two ST steps.

sounds ($GT = 0.16/T^2$), whereas elsewhere, it is twice as high ($GT = 0.32/T^2$), resulting in lesser sensitivity to pitch glide. The DGT is set to 30 ST/s, and the minimum duration of a tonal segment to 0.035 s.

The stylization results in a representation of the audible pitch events in an utterance, which is less complex than the acoustic data. The obtained stylization is shown as a bold line in all figures in this chapter.

7.2.2.4 Differences between implementations of the model The tonal perception model of d'Alessandro and Mertens (1995, section 2.4) includes a short-term integration of F0, which models the limitation of the auditory system to follow rapid changes in pitch. This phenomenon was observed in a study on the perception of vibrato by d'Alessandro and Castellengo (1994), who estimate the perceived pitch using a time-windowed average pitch model. This F0 integration is not incorporated in the Prosogram, because it appears that for very steep F0 falls, pitch intervals and slopes exceed those of vibrato. However, Prosogram uses another type of F0 integration. For F0 variations below the GT, the perceived pitch is approximated by the median F0.

The Prosogram implementation provides more segmentation types than the model of 1995, enabling a more accurate location of the syllabic nucleus when phoneme and syllable alignment are available. In addition, the perceptual thresholds may be adjusted for continuous speech while being sensitive to pauses (see section 7.2.2.3).

7.2.3 Validation through Resynthesis

The stylization based on tonal perception was tested in an experiment (d'Alessandro and Mertens 1995) using TD pitch synchronous overlap and TD-PSOLA (Time Domain Pitch Synchronous Overlap-and-Add) resynthesized versions of sentences, using either the original F0 contour (stimulus V1) or the stylized F0 contour, for three threshold configurations (stimuli V2, V3, V4). The threshold values were chosen such that the stylizations ranged from narrow (V2), to intermediate (V3), to coarse (V4). The thresholds of V2 are those observed in experiments using isolated pure tones ($GT = 0.16/T^2$ and DGT = 20 ST/s). In V3, the GT is twice that of V2. V4 uses very high thresholds ($GT = 0.64/T^2$ and DGT = 60 ST/s). The subjects listened to sentence pairs, containing one stimulus with the original F0 contour (V1) and one with a stylized contour (V2, V3, V4), in randomized order. They had to judge whether the sentences were identical, so the same-different paradigm is used. For untrained listeners, it is very difficult to judge intonation independently from other aspects of speech. The percentage of correct responses is expected to be very high for stimulus pairs V1-V1 and V1-V4 because the former pair is identical and the latter is clearly different as a result of coarse stylization. For V1-V2, the score is expected to be close to that for V1-V1, and for V1-V3, the score is expected to be slightly lower than that for V1-V2. These hypotheses were confirmed by the results. A stylized contour using thresholds observed for isolated sounds (V2 stimuli) is hardly distinguishable from the original. For other threshold settings, the proportion of items judged different correlates with threshold magnitude. Still, the impact of threshold magnitude is moderate and smaller than the role played by syllable segmentation. There appear to be rather important differences between the scores of individual subjects. Also, some part of the error is explained by stimulus differences introduced by resynthesis itself.

Mertens, Beaugendre, and d'Alessandro (1997) report a similar experiment comparing two stylization approaches: the automatic stylization based on tonal perception (ATS) and the manual close-copy stylization (CC), in which the F0 data is manually

approximated by the standardized pitch movements of the IPO model ('t Hart, Collier, and Cohen 1990). Stimuli were prepared using Linear Predictive Coding (LPC) synthesis. The subjects judged whether a stimulus pair was identical or not. A pair consisted of either two stimuli with the original F0 contour (V1) but resynthesized, or one V1 stimulus and one stimulus with the stylized contour obtained with either the ATS (V2A) or the CC (V2B) model. The results indicate that 93 percent of the V1-V1 pairs are judged identical (although 100 percent were identical), whereas V1-V2A and V1-V2B give, respectively, 90 percent and 88 percent identical responses. This shows that resynthesized stimuli of both models are very close, and on average, the ATS model performed slightly better than the CC model did. This may be explained by the fact that the ATS stylization results in a larger number of line segments.

7.3 Prosodic Features Obtained during Stylization

7.3.1 Prosodic Features of Individual Syllables

During stylization, a wealth of prosodic features are computed for each syllable (more precisely for the syllabic nucleus). They are available for statistical analysis (as illustrated by the *prosodic profile*; see Mertens 2020, section 5.3) or further processing, as discussed in sections 7.4 and 7.5. Pitch-related features for each syllable include start, end, maximum, minimum, mean, and median pitch; cumulated upward pitch intervals; cumulated downward intervals; pitch trajectory (sum of absolute upward and downward movements); and pitch interval between the end of the preceding syllable and the start of the current one. All pitch intervals are expressed in ST. Pitch values normalized for the global pitch range of the individual speaker are also available (see section 7.3.2.1). Another feature is the dynamic nature of the syllable (whether it contains a glissando). Duration measures include nucleus duration, as well as vowel, syllable, and rhyme duration, for corpora including phonetic and syllable alignment. Intensity-related measures are peak intensity (measured in dB) and peak loudness (measured from the excitation spectrum). Pause length is available too (see section 7.3.2.2).

A second set of derived, context-related prosodic features includes feature prominence values. A syllable is prominent for a given feature such as duration, loudness, or pitch, when it stands out from the adjacent syllables. This is quantified as the distance between the syllable's value for feature f (e.g., duration) and the mean of the corresponding values in the context. The latter typically consists of two preceding syllables and one following syllable but may be selected otherwise. The context is restricted to a time window (0.5 s on either side of the target syllable) and bounded by the presence of a pause. (The distance is expressed as a proportion or a difference, depending on the linear or logarithmic scale of the feature values.)

7.3.2 Prosodic Features for Sequences of Syllables

Some speakers use a narrow pitch range, others a wide one. Pitch movements may be frequent or rare. Speech may be fast or slow. The prosody of an individual speaker, a speaking style, or a speech fragment is characterized by global prosodic attributes, such as pitch range, speech rate, melodic variability, and isochrony. These attributes are easily derived from the primary prosodic features already mentioned.

7.3.2.1 Pitch range and its normalization The term *pitch range* is often used ambiguously, indicating either the vocal range (tessitura, global pitch span) of a speaker, the register (high, modal, low) used by a speaker (Rietveld and Vermillion 2003), the local

pitch span of a stretch of speech, or even the excursion size of a single pitch movement (Gussenhoven 2004). The pitch range of a speaker may be locally reduced or widened; its central value may be locally raised or lowered (as for register change).

Both global and local pitch range may be characterized by two variables (Ladd 1996, 260–261; 2008, 198): overall level and span. *Overall level* (or *key*, in the terminology of Hirst 2011, 71) is often approximated by the mean or median F0 of the speaker. The span designates the F0 range used by the speaker, to a first approximation the distance between maximum and minimum pitch. Some speakers use a wide span, others a narrow one. The span may be expressed in various ways, but it is usually specified in ST, that is, on a logarithmic scale. As Ladd (2008, 198) observes, overall level and span are often conflated because changes of key and changes of span may be hard to distinguish: raising a high tone produces the same effect as widening the span.

Ladd (1996, 267–269; 2008) examines several quantitative models of pitch range. According to Rietveld and Vermillion (2003), the F0 of low-pitch targets forms the strongest cue to perceived register (better than mean pitch, median pitch, or mid value), and the pitch interval (in ST) between low and high targets provides an estimate of perceived register width. De Looze and Hirst (2010) measure pitch range using the pitch targets obtained by a Momel stylization and propose an algorithm to detect changes in local pitch range.

Within the Prosogram tool, the global pitch range of a speaker is estimated on the basis of the distribution of representative pitch values in the corpus. To exclude pitch detection errors occurring at the transition between a voiceless consonant and a vowel or a glide, pitch data is limited to the sonorant part of the syllable, obtained using the segmentation into syllabic nuclei (section 7.2.2.1). Moreover, pitch discontinuities, often related to octave jumps, are located, and among the parts with continuous pitch, the one nearest to the median F0 of the speaker is selected. Finally, clear outliers (more 18 ST off the median F0 of the speaker) are discarded. To deal with intrasyllabic pitch movements appropriately, two pitch values per syllable are used: the maximum and minimum pitch, after stylization. (For syllables with level pitch, these two values are identical.) The bottom and top of the global pitch range are identified as the second and ninety-eighth percentiles of this distribution.

To compare intonation contours pronounced by different speakers, independently of their pitch range, a normalization of pitch values may be obtained. Normalized pitch contours not only show a high degree of interspeaker agreement, but also greater similarity between pitch contours pronounced by the same speaker in different paralinguistic conditions. Various types of normalization have been proposed, such as the *z*-score of F0, in either hertz or semitones (Jassem and Kudela-Dobrogowska 1980; Poiré and Kaminskaia 2004; de Moraes and Rilliard 2014). Bardiaux and Mertens (2014) instead normalize pitch contours by scaling the stylized pitch values (in ST) by the speaker's pitch range (measured as already described), such that the bottom and top correspond to 0 and 100 percent of the pitch range, respectively.

7.3.2.2 Other prosodic attributes Prosodic features obtained using Prosogram may be used to analyze and detect prosodic attributes and events such as prominence (Goldman, Avanzi, et al. 2007; Christodoulides and Avanzi 2014), prosodic boundaries (Mertens and Simon 2013; Christodoulides and Simon 2015), and turn-taking cues (Heldner and Włodarczak 2015) or to characterize speech styles (Goldman, Auchlin, et al. 2007). They have been used in various other domains, including pathological speech and music research (the relation between speech prosody and music; Patel,

Iversen, and Rosenberg 2006). Because this chapter focuses on pitch stylization and intonation transcription, these aspects will not be discussed here.

7.4 Automatic Symbolic Transcription of Pitch Level and Movement

The pitch stylization described in the previous sections illustrates the mapping between the acoustic and perceptual manifestations of prosody and provides a representation of auditory pitch events, which, although associated with individual syllables, is still continuous in nature. Linguistic analyses of prosody, in contrast, use discrete elements to represent distinctive pitch contrasts in speech. Depending on the linguistic model, these elements will be pitch levels, pitch movements (seen as primitives), tones, or "pitch accents" (Bolinger 1958). They are associated with particular positions in the speech chain, such as the syllable carrying word stress, the boundaries of a syllable sequence, or a boundary of some kind. Given our bottom-up perspective, the question arises whether it is possible to obtain such a discrete linguistic representation starting from the auditory representation.

Most linguistic analyses of prosody include the notion of stress or a related attribute, be it word stress, sentence stress, nuclear tone, pitch accent, or prominence. This prosodic attribute of stress is not available in the stylization described so far. If stress is marked by acoustic features, then it should be possible to detect it using such features. If stress is ultimately a structural property (Gussenhoven 2004), then stress location may be obtained from other sources such as lexical information and syntactic structure. The latter approach is discussed in detail in section 7.5.

Independent of the presence of stress, the categorization of syllabic pitch contours may still be based on principles common to many intonation languages. This section discusses the nature of such principles and proposes a system to categorize pitch contours into discrete elements, as described in Mertens (2013, 2014, 2019).

The ambitious goal of automatic transcription of prosody according to a linguistic annotation scheme may be subdivided into two major parts. The first part provides a generic transcription of pitch levels and movements associated with individual syllables. The second part identifies positions in prosodic structure, mainly stress and boundaries. Both aspects need to be combined according to the prosodic grammar of the language in question, by interpreting pitch levels and movements in terms of these structural positions.

7.4.1 A Generic Transcription of Pitch Levels and Movements

Pitch variation is produced in the larynx; the biophysical properties of the larynx and vocal folds constrain the pitch variations that may occur in speech: maximum rate of pitch change, excursion size, and so forth (Xu 2005). The possible melodic distinctions are also constrained by the limited resolving power of the auditory system ('t Hart, Collier, and Cohen 1990, 74). For speakers of intonation languages (without hearing impairment), the auditory capabilities are fairly similar. So from the perspectives of speech production and perception, there are no fundamental differences between intonation languages, as far as possible pitch movements are concerned; the constraints on pitch variation are basically the same. It seems feasible then to look for a generic transcription.

Languages will differ in the way pitch variations are combined into larger structures, such as pitch contours, or associated with other prosodic or segmental features, such as lexical stress and phrase boundaries. But in the end, every contour is a concatenation

of pitch events occurring on successive syllables. The question then becomes how such contours should be characterized.

Ladd (2008) distinguishes two major approaches to pitch contour characterization, a distinction reminiscent of the levels versus configurations debate (Bolinger 1951). The first view describes pitch contours as configurations, that is, as a sequence of components defined relative to what precedes. Consider a contour consisting of low unstressed syllables, followed by a large high fall on the stressed syllable. Such a configuration specifies the location (stressed syllable, unstressed syllable, boundary), the direction of a movement (rise, fall, level), its melodic interval (large, small), as well as the pitch levels resulting from such intervals (high, low). This approach assumes that a given configuration can occur at various places in the speaker's pitch range without affecting contour identity or function. The second view factors out pitch variation related to the individual speaker or paralinguistic aspects (e.g., bored speech, emotional speech) in order to normalize the tonal space of a speaker (Ladd 2008, 193) into a small number of pitch levels. This approach is exemplified by Pike (1945).

Many prosodic models combine ingredients of both approaches. Autosegmental models, for instance, combine a drastic pitch range normalization (by assuming only two basic pitch levels: low and high) with configurational specifications, describing pitch movements as sequences of pitch targets. In such an approach, pitch contours are characterized as sequences of pitch targets associated with positions defined relative to word stress or prosodic boundaries.

Most intonation models, both configuration- and level-based, refer to global pitch range, although often implicitly. The notion of bottom pitch is a clear example of this. As Ladd (1996, 267) notes, "the *bottom of the speaking range* is a fairly constant feature in an individual's voice." This view is found also in Ladd (2008, 203) and 't Hart (1998, 100). As can be readily observed, listeners are able to detect a speaker's bottom pitch on the basis of phonation characteristics (Honorof and Whalen 2005). As the pitch approaches the bottom of the vocal range, the vibration frequency moves away from the characteristic frequency of the vocal fold vibration; as a result, this vibration may become irregular, as is the case for creaky voice or breathy voice. Creak is not restricted to low-pitched phonation, but its association may be conventional in some language variants or for some speakers.

These considerations regarding pitch range and contour characterization are the basis of the design of a generic transcription system of pitch variation, called Polytonia, which is presented in detail in Mertens (2014). The pitch contour is characterized as a configuration of pitch levels and movements; these levels are defined relative to one another and relative to global pitch range. An intrasyllabic pitch movement is characterized by its starting pitch level, its direction, and its size.

7.4.2 The Proposed Symbolic Tonal Annotation

The tonal transcription specifies pitch level, pitch movement direction, and size for simple or compound movements in a transparent and analytical way. An illustration is given in the second tier of the lower part of figure 7.1, where the symbols L, H, HR, and HF, respectively, indicate a low pitch level, a high pitch level, a large rise starting from a high pitch level, and a large fall starting from a high pitch level. More generally, for each syllable, the transcription indicates the pitch level at vowel onset and, when present, the pitch movement in the rhyme. Complex pitch movements require two or more movement codes, for example, LRF indicates a rise-fall movement starting from a low pitch.

7.4.2.1 Pitch levels The Polytonia annotation defines pitch levels in two ways: locally, as in relative to the context, and globally, as in relative to the speaker's total pitch range. Pitch levels low (L), mid (M), and high (H) are defined locally, relative to one another. A sufficiently large pitch change or movement between a given syllable and some other syllable in the context triggers a change of pitch level. The global interpretation of pitch level, relative to the speaker's pitch range, results in levels top (T) and bottom (B). These additional levels can be explained by their specialized function in speech. In many languages, the bottom pitch marks the end of a maximal prosodic unit or the end of a speaker turn; similarly, the top level is mostly reserved for emotional or emphatic speech portions.

Given this relative successive definition of pitch levels, syllables with the same pitch level code but occurring at different points in the utterance may differ considerably as to their actual F0, provided there are local pitch changes explaining these differences. The overall trend of decreasing pitch, frequently observed in read speech and usually called *declination*, is accounted for in the same way.

7.4.2.2 Pitch intervals Pitch range varies considerably between speakers and speech styles. Hence, pitch interval categories should be expressed relative to pitch range.

The number of pitch interval categories depends on the descriptive model. Many models distinguish between large and small pitch intervals, where the latter typically occur in downstepping or upstepping (IPO model: 't Hart, Collier, and Cohen 1990; INTSINT [International Transcription System for Intonation] model: Hirst and Di Cristo 1998; rhythm and pitch model: Dilley et al. 2006).

The Polytonia transcription system distinguishes two pitch interval sizes: large and small. The size of the small pitch interval is set to 3 ST, which is slightly larger than intrinsic microprosodic variations (Rossi et al. 1981). The large interval is set relative to the observed pitch range (see Mertens 2014, section 5.5, table 2).

7.4.2.3 Intrasyllabic pitch movements There are basically two ways to transcribe a pitch movement within a syllable: as a sequence of pitch levels or as a combination of a pitch level and movement type. For instance, a large rise starting on a low pitch level and ending on a high pitch level will be transcribed respectively as LH or LR, where R represents a large rise. The first approach is common in autosegmental models, which indicate one or more levels (tones) for a syllable and use only two levels (L and H) in combination with downstep and upstep. The limitations of this approach are discussed in Mertens (2014, section 3.2). The second approach is found in some American structuralism works (e.g., Smalley 1964) and is reminiscent of the characterization of contours in the British school (e.g., Crystal 1969; Cruttenden 1986). It allows for a concise and readable representation indicating the pitch level at the start of a syllable, followed by any successive pitch movements in that syllable.

7.4.2.4 Symbol inventory The Polytonia transcription indicates four tonal aspects for each syllable: its starting pitch level, the presence of a simple or compound pitch movement, the shape of each movement (level, rising, or falling), and its size (large or small). All this information can be conveyed by a fairly small set of symbols consisting of five symbols for pitch level (L, H, M, T, and B) and five symbols far large and small pitch movements: R (large rise), F (large fall), r (small rise), f (small fall), and _ (flat). Compound movements use a sequence of the movement symbols, possibly including a level part: for example, RF (rise-fall), _R (level-rise), R_ (rise-level). Although the annotation

can represent compound intrasyllabic pitch movements of any complexity, such movements are fairly rare in intonation languages, even in spontaneous speech.

Provided F0 is detected for a given syllable, its pitch movement is always identified, because it can be interpreted on the basis of the stylized pitch in the syllable itself and the pitch range. Pitch level, however, may not be detected for a given syllable. In such cases the pitch movement is shown without pitch level—for instance, R and RF.

For conciseness, a shorthand notation is used. When pitch movement is level and simple, the symbol indicating the level movement is skipped: H_ is simplified to H, M_ is noted M, whereas H_R and HR_ are noted as such (in this example, the two forms distinguish a late rise from an early one). A level movement with missing pitch level, which initially appears as _, is skipped altogether.

When the pitch level starts from L, M, or H but reaches B or T within the same syllable, this is also indicated—for instance LF,B (large low fall to bottom), Lf,B (small low fall to bottom), and HF,B (large high fall to bottom). In many languages, an intrasyllabic fall ending at the bottom level acts as a terminal boundary in the same way as a flat syllable starting at the bottom level. A terminal boundary ends the largest prosodic domain. Most descriptions of prosody observe that declarative utterances end at the bottom pitch level, suggesting that the function of the bottom level may be common to a large number of languages. Similarly, a syllable that rises to the top level produces the same effect as one starting at that level. The algorithm for the automatic transcription of pitch levels and movements is described in detail in Mertens (2014, sections 5.6 and 5.7).

The next section discusses a more advanced prosody transcription, which takes into account stress and prosodic boundaries, makes explicit prosodic structure, and reduces pitch events to a small set of distinctive forms, applying a language-specific intonation grammar.

7.5 Automatic Symbolic Transcription of Intonation

There is more to intonation than pitch variation. Pitch is just one aspect of prosody, next to duration, loudness, rhythm, and so on. In addition, the communicative function of a pitch event depends on where it occurs in the utterance (more precisely, where it occurs in syntactic structure). For instance, depending on its position, a high pitch or a rise may mark focus or the end of a discourse unit. In linguistic analyses of intonation, the relevant positions are generally defined relative to stress and to boundaries.

In articulatory phonetics, a syllable is stressed when it stands out from its context by its increased phonation effort, articulatory strength, by a pitch change, greater loudness, longer duration, or a combination of these features. (For a similar definition, see Ladd 1996, 58.) Difficulties with the definition of stress led to the introduction of related notions, such as accent, pitch accent (Bolinger 1958), and prominence. Most contemporary intonation grammars refer to the notion of stress or to prominence.

Traditionally, a distinction is made between word stress and sentence stress. *Word stress* (also known as lexical stress; Garde 1968) indicates the syllable within a lexical item that will be stressed when the word is pronounced by itself. Out-of-context pronunciation reveals the virtual stress position within the lexical item. Word stress is a language-dependent property: compare English "photography" and French "photographie." *Sentence stress*, in contrast, indicates a syllable that in an utterance's pronunciation is actually stressed, hence is marked by some of the cues mentioned herein. In utterances of several words, some syllables carrying word stress will be stressed and others unstressed.

Prosodic structure designates an organization of the speech chain, which identifies structural positions, determined by the presence of stress or prosodic boundaries. These

positions are essential for intonation description because the function of a pitch event may be ambiguous when its position is unspecified. Given this underlying structure, the intonation of an utterance is a concatenation of one or more stressed syllables, prosodic boundaries, and unstressed portions between them. Particular sequences form a prosodic unit, such as an "intonation phrase." Phonological models usually distinguish several layers of hierarchical structuring with their corresponding prosodic units. (Empirical observation from manual and automatic transcription of prosody in colloquial speech suggests that some of these layers are highly speculative.)

Given the fact that prosodic structure refers to stress and indirectly to word stress, which is language-dependent, an intonation model will inevitably be specific to a given language and so will be the corresponding representation and the automatic procedure to compute it.

Having clarified the notions of prosodic structure, stress, and intonation units, we now turn to automatic decoding of this structure. To obtain a comprehensive transcription of intonation, specifying structural positions in the utterance as well as the pitch events occurring for them, information is needed about the location of stressed syllables and prosodic unit boundaries.

In what follows, I illustrate the automatic computation of such a transcription for French, first however summarizing French prosodic structure.

Two layers of organization may be identified in French intonation (Mertens 2006; Mertens et al. 2001). On a first layer, the stress group consists of a lexical item carrying word stress and the adjacent words (mostly grammatical words) that syntactically depend on it. So stress groups may be derived from word stress and syntactic dependency relations between words. Each stress group contains a syllable with virtual stress. On a hierarchically higher layer, the intonation group consists of a sequence of one or more stress groups, in which the last carries sentence stress, whereas the other virtual stress positions remain unstressed. For the phonological characterization of distinctive pitch contours observed on intonation groups in French, three structural positions are needed: the final stressed syllable, the penultimate, and the syllable carrying initial stress. The latter position is optional within the intonation group. When present, it usually coincides with the initial syllable of a lexical item, rarely with the second, and never with final stress.

Speech corpora that provide (temporally aligned) annotations for word forms, part of speech, sounds, syllables, stress (or prominence), and a pitch labeling (such as Polytonia) allow for the computation of pitch contours, specifying their structural positions and the pitch events (pitch level, pitch movement, or tone) associated with these positions. The resulting annotation is illustrated in figure 7.6, which shows the input annotation tiers just mentioned and the output: the structural positions in the pitch contour with their respective labels. The same information could be represented by association lines connecting tones to their position in the segmental layer, in combination with stress and boundary marks, as is common in autosegmental approaches. The structural positions are indicated as follows. Stress is marked by the IPA stress mark preceding the pitch level; for instance ˈH indicates a syllable with final stress, on a high pitch level. Initial stress is differentiated from final stress by the suffix *i* appended to the pitch symbol—for instance ˈMi. The penultimate position is noted by the suffix *p*, as in Lp. Finally, X is used when some input information is missing (such as the pitch level on the penultimate of *simplifie*) or for syllables with a special status (such as the schwa following the stressed syllable in *courbe*).

The transcription of a longer fragment appears in figure 7.7, with just two annotation layers: word form and intonation transcription.

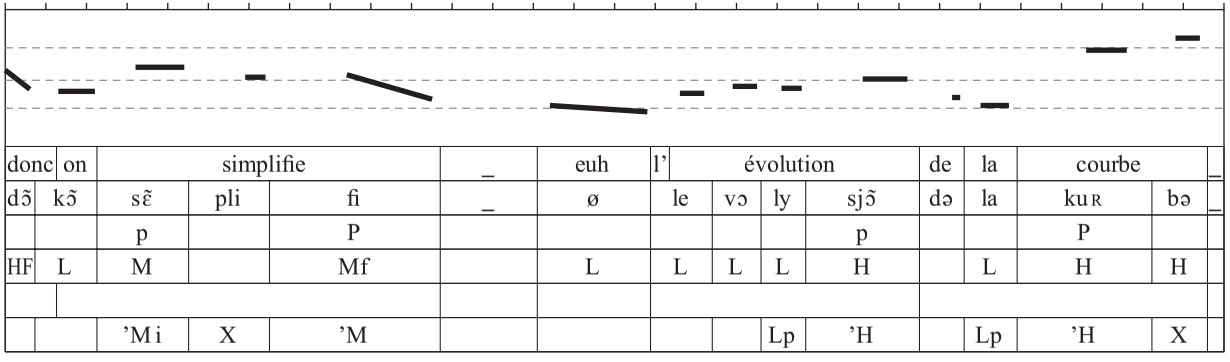


Figure 7.6
Input and output for the automatic transcription of intonation contours for a short sample of the C-Prom corpus (Avanzi et al. 2010) of French, file cnf-be, which records an academic presentation by a female speaker. The utterance is "Donc on simplifie euh l'évolution de la courbe" (So we simplify the shape of the curve). The annotation tiers indicate (i) word form, (ii) syllable, (iii) prominence (*P* and *p* indicate strong and weak prominence, respectively), (iv) pitch level and movement (obtained by Polytonia), (v) the intonation groups, and (vi) structural positions within the pitch contour with their respective pitch labels. In tier 6, final stress is marked by an IPA stress mark preceding the pitch symbol, for example, ˈH; initial stress is also indicated by a stress diacritic and an *i* suffix appended to the pitch symbol, for example, ˈMi; the penultimate syllable is noted by the suffix *p*, for example, Lp. The *X* symbol indicates missing information or a syllable with a special status.

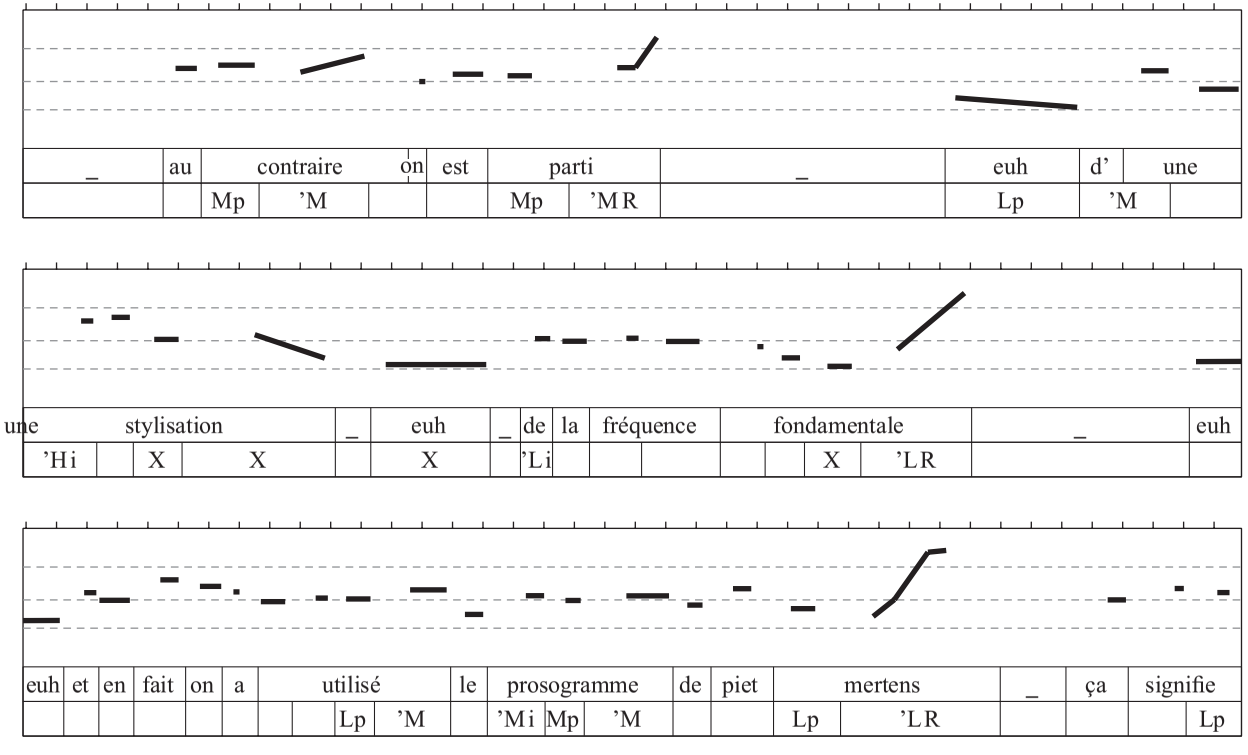


Figure 7.7
Automatic transcription of intonation units and contours for a fragment of the C-Prom corpus, recording "cnf-be," pronounced by a female speaker. The annotation tiers indicate word form and intonation contour transcription using the conventions of figure 7.6. The passage translates to: "On the contrary, we started from a stylization of fundamental frequency. And in fact we used the Prosogram of Piet Mertens. This means . . ."

The system for automatic transcription of intonation in French, illustrated in figures 7.6 and 7.7, was elaborated in collaboration with Sandrine Brognaux (2015) and is called ToPPos-Fr, for "tones and prosodic positions in French." It uses several inputs (information sources), which may be either specified explicitly by annotation tiers or computed by dedicated modules on the basis of additional information sources. For instance, prominence information may be provided by the corpus annotation (as is the case for the C-Prom corpus used in the figures) or by a prominence detection module using acoustic cues (Brognaux 2015). Similarly, stress groups may be determined using part-of-speech information, together with local syntactic relations (or, alternatively, from the enumeration of part-of-speech sequences that form stress groups). Two further steps follow: one for the identification of structural positions and the other for mapping pitch levels and movements onto the set of tonal categories that the intonation model uses. The next two paragraphs provide more details on these steps.

For identifying the three structural positions of final, penultimate, and initial stress, a set of rules is applied to each syllable. Final stress position is assigned to a prominent syllable carrying word stress, that is, when it is the final syllable of a stress group and that syllable does not contain a schwa vowel. Final stress position then forms the right boundary of the corresponding intonation group. An optional schwa may occur after final stress within the same word (as is the case for the word *courbe* in the example in figure 7.6); this schwa then forms the last syllable of the intonation group but is not itself in final stress position. Initial stress position is inferred when a prominent syllable is word-initial but not in final stress position. The syllable preceding a final stress is labeled "penultimate." Additional rules deal with special cases. For instance, the last syllable of a stress group that is followed by a pause (or by a schwa and a pause) and reaches the bottom pitch level is considered to be in final stress position, whether it is prominent or not.

As a final step, pitch levels and movements (as identified by Polytonia) are mapped onto the categories used in the intonation model targeted by the transcription system. Tier 2 of figure 7.7 illustrates one such model or transcription convention: ToPPos-Fr. More generally, a mapping may entail a reduction of the number of retained pitch levels, a simplification of compound movements to simple ones, or a conversion of pitch movements to sequences of pitch levels. Moving categorization even further, the fairly large set of observed pitch events may be mapped onto a small set of functional categories such as major continuation, minor continuation, terminal contour, question contour, and focus, and so forth. Whatever transcription convention is selected, the system architecture remains unchanged; only the mapping rules differ.

The system for automatic transcription of intonation contours in French outlined in this section simulates perceptual and cognitive processing. When we decode the intonation of an utterance, we combine multiple sources of information, both perceptual and grammatical (our knowledge of the language grammar). This interaction process is simulated by a rule-based module.

7.6 Conclusion

Linguistic analysis of intonation aims at identifying the distinctive prosodic forms of spoken language. It studies their alignment with segmental structures, be it syllables, lexical items, syntactic structures, or discourse units. It also investigates the meaning of the primitive forms, both taken separately and in interaction with other linguistic

structures (syntactic constructions, discourse structure). This chapter is limited to the formal aspects of prosody: the auditory forms used by intonation.

The central assumptions of our approach may be summarized as follows.

Perceptual processing in the auditory system affects the way listeners interpret F0 variations in speech. Spectral instability introduces a segmentation into the speech signal and, hence, in the evolution of its acoustic prosodic parameters. The resulting representation in auditory memory consists of a chain of elements that correspond to local maxima of sonority with relative spectral stability and correspond roughly to the syllabic nuclei.

The proposed type of pitch stylization simulates the impact of this segmentation on the perception of pitch in speech. In the stylized contour, a syllable's pitch is either static (flat) or dynamic (when it includes an audible pitch glide), and the evolution of pitch during speech is characterized by the melodic intervals that separate successive syllables or occur within individual syllables. This approach aspires to modeling psychological reality and cognition while at the same time providing a model of acoustic F0 contours that has been applied successfully in resynthesis, text-to-speech synthesis, and intonation transcription.

In linguistic decoding of speech prosody by humans, pitch variations are normalized relative to the global pitch range of the individual speaker. The size of pitch steps and movements is evaluated relative to the speaker's pitch span. The bottom and top of his pitch range are partially identified from phonation properties.

In many transcriptions and models of prosody used in linguistics, the evolution of pitch is represented by pitch levels and pitch movements. These may be computed automatically, starting from the pitch stylization based on tonal perception, as is illustrated by the Polytonia program (section 7.4).

Comprehensive linguistic models of intonation postulate a prosodic structure, which takes into account stress and boundaries, defines prosodic units (such as the intonation phrase), and characterizes pitch contours as sequences of pitch events (tones) associated with positions in prosodic structure. The actual grammar will differ from one language to the next. By combining multiple information sources, such intonation contour specifications may be derived automatically, as illustrated by the program set out in section 7.5.

The latter application provides a specification of intonation matching the requirements of a so-called *phonological intonation model*. It posits a set of abstract, symbolic primitives (pitch levels and movements) that combine with positions (stress, prestress, emphatic stress, boundaries, and so on) in prosodic structure to generate well-formed intonation contours, as well as their acceptable phonetic instantiations (by synchronizing target points with syllable structure).

Pitch levels and movements may be viewed as the phonemes of prosody, defined relative to one another and without meaning of their own. An intonation morpheme is formed when one or more phonemes are associated with a given position in prosodic structure, such as stress. Examples of such morphemes in French are (using the notation of section 7.5) 'H, 'M, 'HF, 'LR, 'Hi, 'B, and Hp. Such intonation morphemes do have a meaning and may form minimal pairs. The sequence of morphemes within a prosodic unit forms an intonation contour, which ultimately may be viewed as a syntactic constituent of the prosodic domain. In this view, the concepts of phoneme, morpheme, and phrase in the prosodic domain parallel those of the segmental domain.

Meaning or function of intonation contours is not dealt with in this chapter, which focuses on bottom-up analysis, recognition, and transcription of the prosodic

primitives and their organization into larger units. Propositions for a compositional characterization of intonation meaning in French and for the interface with syntactic constructions are in Mertens (2006). As far as intonation meaning is concerned, formal, objective, reproducible methods appear to be scarce.

The output representations (stylization, symbolic pitch levels and movements, intonation transcription) derived by the model have been used as input to a text-to-speech system (Mertens et al. 2001). In this respect, our approach is related to the analysis by synthesis approach that Hirst (2011) proposed.

In his critical review of prosody research methodology, Xu (2011) distinguishes several approaches: analysis by transcription, analysis by introspection, analysis by hypothesis testing, and analysis by modeling. Our approach belongs mainly to the last type. Xu (2011, section 5) writes: "Potentially the most rigorous test of our understanding of prosody, especially in terms of predictive knowledge, is computational modeling. . . . Its importance has in general not been duly recognized." Because it combines acoustic, perceptual, and linguistic representations of prosody, our approach may also be characterized as an integrated model of prosody. It provides a multidisciplinary synthesis of methods and insights from auditory (manual) transcription of intonation, articulatory phonetics, acoustic phonetics, psychoacoustics, phonology, linguistics, speech processing, computing, and natural language processing.

Independent of such methodological considerations, Prosogram provides a tool that enables researchers to easily obtain a quantified, objective, and perceptually motivated representation of prosody in continuous speech (Mertens 2020). It is our hope that the availability of this tool will encourage researchers not to take prosodic theories for granted, but to discover the richness of prosody, as it emerges from large-scale corpora of spontaneous speech.

References

Avanzi, M., A. C. Simon, J.-Ph. Goldman, and A. Auchlin. 2010. "C-PROM: An Annotated Corpus for French Prominence Studies." In *Proceedings of Prosodic Prominence: Perceptual and Automatic Identification, Speech Prosody 2010 Workshop*. https://www.isca-speech.org/archive/sp2010/papers/sp10_2005.pdf.

Baken, R. J., and R. F. Orlikoff. 2000. *Clinical Measurement of Speech and Voice*. San Diego: Singular Publishing.

Bardiaux, A., and P. Mertens, 2014 "Normalisation des contours intonatifs et étude de la variation régionale en français." *Nouveaux cahiers de linguistique française* 31:273–284.

Barnes, J., A. Brugos, N. Veilleux, and S. Shattuck-Hufnagel. 2014. "Segmental Influences on the Perception of Pitch Accent Scaling in English." In *Proceedings of Speech Prosody 7*, edited by N. Campbell, D. Gibbon, and D. Hirst, 1125–1129. https://www.isca-speech.org/archive/SpeechProsody_2014/pdfs/219.pdf.

Boersma, P., and D. Weenink. 2012. *Praat: Doing Phonetics by Computer*. Version 5.4.18. Computer program. http://www.praat.org/.

Bolinger, D. L. 1951. "Intonation: Levels vs. Configurations." *Word* 7:199–210.

Bolinger, D. L. 1958. "A Theory of Pitch Accent in English." *Word* 14:109–149.

Brognaux, S. 2015. "Expressive Speech Synthesis: Research and System Design with Hidden Markov Models." PhD diss., Université catholique de Louvain and Université de Mons.

Christodoulides, G., and M. Avanzi. 2014. "An Evaluation of Machine Learning Methods for Prominence Detection in French." In *Proceedings of Interspeech*. https://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_0116.pdf.

Christodoulides, G., and A. C. Simon. 2015. "Exploring Acoustic and Syntactic Cues to Prosodic Boundaries in French: A Multi-Genre Corpus Study." In *Proceedings of the International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/proceedings.htlm.

Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

d'Alessandro, C., and M. Castellengo. 1994. "The Pitch of Short-Duration Vibrato Tones." *Journal of the Acoustical Society of America* 95:1617–1630.

d'Alessandro, C., and P. Mertens. 1995. "Automatic Pitch Contour Stylization Using a Model of Tonal Perception." *Computer Speech and Language* 9 (3): 257–288.

De Looze, C., and D. J. Hirst. 2010. "Integrating Changes of Register into Automatic Intonation Analysis." In *Proceedings of the Speech Prosody Conference*. https://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_0116.pdf.

De Looze, C., and D. Hirst. 2014. "The OMe (Octave-Median) Scale: A Natural Scale for Speech Melody." In *Proceedings of the Speech Prosody Conference*.

de Moraes, J. A., and A. Rilliard. 2014. "Illocution, Attitudes and Prosody." In *Spoken Corpora and Linguistic Studies*, edited by T. Raso and H. Mello, 233–270. Amsterdam: John Benjamin.

Dilley, L., M. Breen, E. Gibson, M. Bolivar, and J. Kraemer. 2006. "A Comparison of Inter-Coder Reliability for Two Systems of Prosodic Transcriptions: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices)." In *Proceedings of the International Conference on Spoken Language Processing*. http://speechlab.cas.msu.edu/PDF/Old Conference Proceedings/Dilley Breen Bolivar Kraemer Gibson 2006.pdf.

Garde, P. 1968. *L'accent*. Paris: Presses Universitaires de France.

Goldman, J.-Ph., A. Auchlin, A. C. Simon, and M. Avanzi. 2007. "Phonostylographe: Un outil de description prosodique. Comparaison du style radiophonique et lu." *Nouveaux Cahiers de Linguistique Française* 28:223–241.

Goldman, J.-P., M. Avanzi, A. Lacheret-Dujour, A. Simon, and A.-C. Auchlin. 2007. "A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French." In *Proceedings of Interspeech*. https://www.isca-speech.org/archive/interspeech_2007/i07_0098.html.

Grice, M. 2006. "Intonation." In *Encyclopedia of Language and Linguistics*, 2nd ed., edited by K. Brown, vol. 5, 778–788. Oxford: Elsevier.

Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Heldner, M., and M. Włodarczak. 2015. "Pitch Slope and End Point as Turn-Taking Cues in Swedish." In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0307.pdf.

Hermes, D. J. 1987. "Vowel-Onset Detection." *IPO-Annual Progress Report* 22:15–24.

Hermes, D. J. 2006. "Stylization of Pitch Contours." In *Methods in Empirical Prosody Research*, edited by S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schließer, 29–61. Berlin: Walter de Gruyter.

Hermes, D. J., and J. C. Van Gestel. 1991. "The Frequency Scale of Speech Intonation." *Journal of the Acoustical Society of America* 90:97–102.

Hirst, D. J. 2011. "The Analysis by Synthesis of Speech Melody: From Data to Models." *Journal of Speech Science* 1 (1): 55–83.

Hirst, D. J., and A. Di Cristo. 1998. "A Survey of Intonation Systems." In *Intonation Systems: A Survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 1–44. Cambridge: Cambridge University Press.

Hirst, D. J., A. Di Cristo, and R. Espesser. 2000. "Levels of Representation and Levels of Analysis for Intonation." In *Prosody: Theory and Experiment*, edited by M. Horne, 51–87. Dordrecht: Kluwer Academic Publishers.

Hirst, D. J., P. Nicolas, and R. Espesser. 1991. "Coding the F0 of a Continuous Text in French: An Experimental Approach." In *Proceedings of the International Congress of Phonetic Sciences*, 234–237. Aix-en-Provence: Université de Provence. Service des publications.

Honorof, D. N., and D. H. Whalen. 2005. "Perception of Pitch Location within a Speaker's F0 Range." *Journal of the Acoustical Society of America* 117 (41): 2193–2200.

House, D. 1990. *Tonal Perception in Speech*. Lund, Sweden: Lund University Press.

House, D. 1995. "The Influence of Silence on Perceiving the Preceding Tonal Contour." *Proceedings of the International Congress on Phonetic Sciences 13* vol. 1, 122–125. https://www.internationalphoneticassociation.org/icphs/icphs1995.

House, D. 1996. "Differential Perception of Tonal Contours through the Syllable." In *Proceedings of the International Conference of Spoken Language Processing*, 2048–2051. http://www.asel.udel.edu/icslp/cdrom/vol4/064/a064.pdf.

Jassem, W., and K. Kudela-Dobrogowska. 1980. "Speaker-Independent Intonation Curves." In *The Melody of Language: Intonation and Prosody*, edited by L. R. Waugh and C. H. van Schooneveld, 135–148. Baltimore: University Park Press.

Lacheret-Dujour, A., S. Kahane, and P. Pietrandrea, eds. 2019. *Rhapsodie. A Prosodic and Syntactic Treebank for Spoken French*. Amsterdam: Benjamins.

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

Ladd, D. R. 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Mertens, P. 2004a. "Un outil pour la transcription de la prosodie dans les corpus oraux." *Traitement automatique des langues* 45 (2): 109–130.

Mertens, P. 2004b. "The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model." In *Proceedings of Speech Prosody 2004*, edited by B. Bel and I. Marlien. Aix-en-Provence: Laboratoire Parole et Langage.

Mertens, P. 2006. "A Predictive Approach to the Analysis of Intonation in Discourse in French." In *Prosody and Syntax*, edited by Y. Kawaguchi, I. Fonagy, and T. Moriguchi, 64–101. Amsterdam: John Benjamins.

Mertens, P. 2013. "Automatic Labelling of Pitch Levels and Pitch Movements in Speech Corpora." In *Proceedings TRASP 2013, Tools and Resources for the Analysis of Speech Prosody*,

edited by B. Bigi and D. Hirst, 42–46. http://www2.lpl-aix.fr/~trasp/Proceedings/TRASP2013_proceedings.pdf.

Mertens, P. 2014. "Polytonia: A System for the Automatic Transcription of Tonal Aspects in Speech Corpora." *Journal of Speech Sciences* 4 (2): 17–57.

Mertens, P. 2019. "From Pitch Stylization to Automatic Tonal Annotation of Speech Corpora." In *Rhapsodie. A Prosodic and Syntactic Treebank for Spoken French*, edited by A. Lacheret-Dujour, S. Kahane, and P. Pietrandrea, 233–250. Amsterdam: Benjamins.

Mertens, P. 2020. *Prosogram User's Guide*. https://sites.google.com/site/prosogram.

Mertens, P., F. Beaugendre, and Ch. d'Alessandro. 1997. "Comparing Approaches to Pitch Contour Stylization for Speech Synthesis." In *Progress in Speech Synthesis*, edited by J. P. H. vanSanten, R. W. Sproat, J. P. Olive, and J. Hirschberg, 347–363. New York: Springer-Verlag.

Mertens, P., J.-Ph. Goldman, É. Wehrli, and A. Gaudinat. 2001. "La synthèse de l'intonation à partir de structures syntaxiques riches." *Traitement Automatique des Langues* 42 (1): 145–192.

Mertens, P., and A. C. Simon. 2013. "Towards Automatic Detection of Prosodic Boundaries in Spoken French." In *Proceedings of the Prosody-Discourse Interface Conference*, edited by P. Mertens and A. C. Simon, 81–87. https://www.arts.kuleuven.be/ling/cohistal/conference/idp2013/documents/proceedings_idp2013.

Nolan, F. 2003 "Intonational Equivalence: An Experimental Evaluation of Pitch Scales." In *Proceedings of the International Congress of Phonetic Sciences*, 771–774. https://www.internationalphoneticassociation.org/icphs/icphs2003.

Patel, A. D., J. R. Iversen, and J. C. Rosenberg. 2006. "Comparing the Rhythm and Melody of Speech and Music: The Case of British English and French." *Journal of the Acoustical Society of America* 119:3034–3047.

Pike, K. L. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Poiré, Fr., and S. Kaminskaia. 2004. "Comparing Intonation of Two Varieties of French Using Normalized F0 Values." In *Proceedings of the Eighth International Conference on Spoken Language Processing*, 1305–1308. https://www.isca-speech.org/archive/archive_papers/interspeech_2004/i04_1305.pdf.

Rietveld, T., and P. Vermillion. 2003. "Cues for Perceived Pitch Register." *Phonetica* 60 (4): 261–272.

Rossi, M. 1971. "Le seuil de glissando ou seuil de perception des variations tonales pour la parole." *Phonetica* 23:1–33.

Rossi, M. 1978. "Interactions of Intensity Glides and Frequency Glissandos." *Language and Speech* 21:384–396.

Rossi, M., A. Di Cristo, D. Hirst, Ph. Martin, and Y. Nishinuma. 1981. *L'intonation. De l'acoustique à la sémantique*. Paris: Klincksieck.

Scheffers, M. T. M. 1988. "Automatic Stylization of F0-Contours." In *Proceedings of the Seventh FASE Symposium*, edited by W. A. Ainsworth and J. N. Holmes, 981–987. Edinburgh: Institute of Acoustics.

Smalley, W. A. 1964. *Manual of Articulatory Phonetics*. New York: Practical Anthropology.

Spaai, G. W. G., A. Storm, A. S. Derksen, D. J. Hermes, and E. F. Gigi. 1993. *An Intonation Meter for Teaching Intonation to Profoundly Deaf Persons*. IPO Manuscript 968. Eindhoven: Technische Universiteit Eindhoven.

Taylor, P. 1994. "The Rise/Fall/Connection Model of Intonation." *Speech Communication* 15 (1/2): 169–186.

Taylor, P. 2000. "Analysis and Synthesis of Intonation Using the Tilt Model." *Journal of the Acoustical Society of America* 107 (3): 1697–1714.

't Hart, J. 1998. "Intonation in Dutch." In *Intonation Systems: A Survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 96–111. Cambridge: Cambridge University Press.

't Hart, J., R. Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.

Xu, Y. 2005. "Speech Melody as Articulatorily Implemented Communicative Functions." *Speech Communication* 46:220–251.

Xu, Y. 2011. "Speech Prosody: A Methodological Review." *Journal of Speech Sciences* 1 (1): 85–115.

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