Journal of Speech Sciences 4(2) 17-57.2014
Available at: www.journalofspeechsciences.org

Polytonia: a system for the automatic transcription of tonal aspects in speech corpora

Piet Mertens
Department of Linguistics, K.U.Leuven

Abstract

This paper first proposes a labeling scheme for tonal aspects of speech and then describes an automatic annotation system using this transcription.

This fine-grained transcription provides labels indicating pitch level and pitch movement of individual syllables. Of the five pitch levels, three (low, mid, high) are defined on the basis of pitch changes in the local context and two (bottom, top) are defined relative to the boundaries of the speaker's global pitch range. For pitch movements, both simple and compound, the transcription indicates direction (rise, fall, level) and size, using size categories (pitch intervals) adjusted relative to the speaker's pitch range.

The automatic tonal annotation system combines several processing steps: segmentation into syllable peaks, pause detection, pitch stylization, pitch range estimation, classification of the intra-syllabic pitch contour, and pitch level assignment. It uses a dedicated and rule-based procedure, which unlike commonly used supervised learning techniques does not require a labeled corpus for training the model.

The paper also includes a preliminary evaluation of the annotation system, for a reference corpus of nearly 14 minutes of spontaneous speech in French and Dutch, in order to quantify the annotation errors. The results, expressed in terms of standard measures of precision, recall, accuracy and F-measure are encouraging. For pitch levels low, mid and high an F-measure between 0.946 and 0.815 is obtained and for pitch movements a value between 0.708 and 1.

Provided additional modules for the detection of prominence and prosodic boundaries, the resulting annotation may serve as an input for a phonological annotation.

Keywords: prosody; transcription; annotation; automatic labeling; pitch range

1 Introduction

It is widely acknowledged that the understanding of prosody is essential for the development of speech technology applications, such as text-to-speech synthesis and human-machine dialog systems, and that progress in this area requires large speech corpora, containing a reliable and objective representation of prosody. Intonation research in phonetics and linguistics, too, would benefit largely from speech *corpus annotations* indicating pitch contours, stress and prosodic boundaries. Yet, prosodically annotated corpora are scarce, for languages other than English. Since the time required for manual annotation is prohibitive, the only alternative is to obtain a prosodic annotation by using an *automatic labeling system*. The design and evaluation of such a system constitutes the topic of this paper. It describes a system for the *automatic transcription* of *tonal attributes* in speech corpora, providing a transcription identifying pitch levels and pitch movements associated with syllables or sequences of syllables. This system will be referred to as the *Polytonia* system.

Several types of information are used by the automatic transcription system: acoustic parameters, the phonetic alignment and the speaker turn annotation. Acoustic parameters include f0, voicing, intensity and duration (of syllabic nuclei). The phonetic alignment provides a segmentation of the signal in speech sounds and derived units, such as the syllable and the syllable rhyme. It should be stressed that no lexical or grammatical information (such as part-of-speech) is used. Our goal is to identify prosodic forms on the basis of acoustic information alone, rather than by using lexical or grammatical information, such as the distinction between function words and content words, or the location of syllables carrying word stress.

Prosodic properties of speech may be characterized at various *levels of observation*. At the acoustic level one measures physical parameters of the speech signal, such as fundamental frequency (f0), sound duration, syllable duration, pause duration, voicing, and intensity. At the perceptual level one tries to establish to what extent acoustic prosodic events are perceived by the average listener and in what way. The level of phonological description, finally, posits a minimal set of distinctive forms, which are assumed to have a function in speech communication. The representations for these levels of observation differ, while being related: an abstract phonological form may correspond to several forms of the perceptual level, the shapes of which differ to some extent, while being related. In the IPO model (‘t Hart et al., 1990), for instance, a first contour with gradual pitch rise occurring during a vowel, and a second contour consisting of a low plateau followed by a fast rise would both be categorized at the phonological level as instances of a “late rise”. All three levels of observation are relevant to the system described in this paper: starting from acoustic measurements, it simulates tonal perception, and applies a further categorization of pitch change, which may be regarded as a preliminary step preceding phonological structuring.

The following paragraphs discuss the choice of a transcription convention for prosody.

Every system for automatic annotation of prosody faces the fundamental question about the very nature of such a transcription. Which aspects of prosody should be indicated and in what way? Which set of symbols should be used in the annotation?

According to some, pitch variation is not perceived categorically, and hence any transcription of pitch using categories would be inappropriate. However, given the various functions of pitch in speech, it is not unlikely that the corresponding perceptual tasks use specialized perception modes: identifying sentence type, identifying focus, interpreting the emotions of a speaker, identifying his regional pronunciation, recognizing the melody a song, and so on. Also, the wealth of intonation models using discrete pitch units (such as tones) suggests that, at some level of representation, categories are meaningful.

Numerous *annotation systems* for prosody have been proposed; an overview is provided by Martin (2009). They may be classified according to various criteria.

A first criterion concerns the *prosodic features indicated* in the notation. While annotation systems such as the IPO model (‘t Hart et al., 1990), INTSINT (Hirst and Di Cristo, 1998, Hirst 2005), the Tilt model (Taylor, 2000), and the PENTA model (Xu, 2005) only represent pitch variations, others also provide information about attributes such as stress, prominence, metrical prominence (Dilley et al., 2006), pitch accent (ToBI; Silverman et al., 1992; Beckman et al., 2005), or boundaries (ToBI), or organise these aspects in a hierarchical manner, where the presence of some attribute (such as accent) implies the presence of a tone, for instance. A second criterion distinguishes *generic* annotations (such as INTSINT), intended for multiple languages, from annotations designed for a particular language (such as American English or German). A third criterion opposes *broad* transcriptions, indicating only distinctive properties, and *narrow* ones, marking aspects relevant for dialectal variation (e.g. the IViE transcription; Grabe et al., 2001), speech style and emotional speech. A last criterion concerns the level of abstraction of the transcription and opposes *acoustic*, *perceptual and phonological* models.

From a linguistic point of view, the interest of phonological representations of prosody lies in the fact that they only represent distinctive intonation phenomena, which are supposed to be related to communicative functions, ignoring contextual, regional, or free variation. This approach results in a fairly small set of symbols. Although the nature and number of these symbols largely vary according to the model, most models indicate the following: syllable stress, a set of tones representing pitch levels and boundaries between prosodic units.

Phonological transcription systems are generally closely linked to a particular phonological theory, such as the autosegmental metrical (AM) framework. For American English, this framework posits only two pitch levels (tones) which may be used in particular positions (on the syllable carrying pitch accent, at the start or at the end of a prosodic unit…) as well as a number of structural units (the intonation phrase, the intermediate phrase...), delimited by prosodic boundaries. Although autosegmental models come in various flavours (for an overview, see Ladd, 1996, 2008 and Grice, 2006), the best known example is the ToBI annotation scheme (Silverman et al., 1992; Beckman et al., 2005). It is widely used for American English, but the principles have also been applied to other languages (Jun, 2005) such as Japanese, Korean, German, Greek, and Dutch, resulting in transcription systems for particular languages. It provides a broad representation of intonation, indicating "pitch accent", and using only two pitch levels, combined with a (small interval) "down-step" movement, and "boundary tones".

Other phonological or phonetic transcription systems may be more fine-grained, identifying more pitch levels, and various types of pitch movements within the syllable. The INTSINT notation ("International transcription system of intonation") (Hirst and Di Cristo, 1998; Hirst, 2005, 2011) was originally intended as a tool for the systematic transcription of intonation patterns in languages, including those for which no phonological analysis of intonation is available. It is based on the inventory of pitch contrasts found in published descriptions of intonation (Hirst 2005, p. 431). INTSINT distinguishes absolute levels (Top, Mid, Bottom), relative levels (Higher, Same, Lower), and iterative relative levels (Up-stepped, Down-stepped). The IViE notation (Grabe et al., 2001) describes the shape and alignment of pitch patterns relative to the location of accented (i.e. stressed) syllables. Consequently it uses larger units, consisting of the accented syllable, the preceding syllable, and the following ones. Campione and Véronis (2001) propose a labeling scheme with 3 major pitch movements (rising, falling, level), emphatic stress, and pauses (medium, long and short). These few examples illustrate the diversity of the approaches used for transcription.

When applying such annotation conventions to corpora of spontaneous speech, two types of *problems* arise. On the one hand, in some cases the inventory of tones and locations seems too small to capture pitch differences which are considered relevant. As a result prosodic contours which are felt to be distinct are nevertheless collapsed in the notation. On the other hand, the proposed structural units sometimes prove hard to keep apart, as shown by the lack of agreement between transcribers for particular forms, in particular for spontaneous speech (Escudero et al., 2012).

To avoid these shortcomings, a *different approach* is followed here, which aims at genericity and descriptive detail and avoids dependency on any phonological theory of prosody. As can be readily observed, the same prosodic phenomena receive different notations in alternative phonological models. Hence, a common representation can only be achieved at a *lower level of abstraction*, which could be either the acoustic parameters or the perceptual representation of prosody, resulting from the processing of acoustic information in the peripheral auditory system and the perceptual system.

The approach followed here aims at a *perceptual representation of pitch phenomena*. It applies a pitch stylization based on a model of *tonal perception* in speech (d'Alessandro and Mertens, 1995; Mertens et al., 1997; Mertens, 2004a, 2004b) to process the acoustic prosodic information and in this way obtain a representation of the audible pitch events in an utterance, which is less complex than the acoustic data itself. In a next step, these stylized pitch events are further categorized into a fairly small set of classes, defined on the basis of their *formal properties only,* and *not by their function* or distinctive nature. This procedure results in a transcription of pitch levels and pitch movements associated with syllables or sequences of syllables. Still, the resulting transcription is more fine-grained than that used in phonological approaches, while being highly readable. Also, the proposed approach is *strictly bottom-up*, which means it is based as much as possible on the information present in the speech signal.

Pitch variations, their shape, excursion, and synchronization, seem similar in most intonation languages. This is not surprising: from the point of view of speech production, there are hardly any differences between languages. Pitch variations are produced in the larynx and the biophysical properties of the speech organ determine its capabilities (Xu, 2005, p. 223). Moreover the auditory capabilities are fairly similar for most speakers of intonation languages. Therefore the perceptual representation makes a good candidate for a language-independent representation of pitch phenomena.

The proposed transcription *does not aim at a phonological analysis* of prosody. It does *not* provide information on stress, prominence, or prosodic boundaries, although it is acknowledged that these aspects play a central role in prosodic structure and would be part of a more comprehensive annotation of prosody. The restriction to pitch-related aspects is a deliberate choice, which will be discussed in detail in section 7. For the time being let us mention that it is motivated by the aim of a prosody annotation which achieves the following goals: it is language-independent, it is independent of a particular phonological model of prosody, it is obtained from acoustic information only, and it takes into account tonal perception. The correlates of stress and boundaries may differ among languages and their treatment largely varies between linguistic models. As a result, every comprehensive annotation of prosody is likely to be restricted to a particular language and a particular model of prosody. Although prosodic structure is not taken into account here, we will argue that information on stress and boundary location, when available, may be combined with the tonal annotation into a comprehensive representation of prosody and that this can be achieved in straightforward manner. We return to this point in section 7.

The annotation which will be proposed here differs from the phonological ones in several ways. It is a *narrow* (i.e. fine-grained) transcription, which represents not only pitch level, but also direction, size and complexity of the pitch movement within a syllable. Pitch level is defined both locally, relative to the pitch levels in the context, and in a global manner, relative to the pitch range of the speaker. The annotation distinguishes two sizes of pitch movements and allows for successive movements within a single syllable: rise-fall, rise-level, level-rise…

The remainder of this paper is organized as follows. Section 2 discusses the characterization of pitch contours in relation to the concept of pitch range. Section 3 describes the proposed tonal annotation, the inventory of symbols indicating pitch level and pitch movements and the grammar for the combination of these symbols. Section 4 provides an overview of existing approaches to automatic transcription of prosody and the criteria for their classification. Section 5 presents a detailed description of the automatic transcription system, focusing on the algorithms used in the successive steps: segmentation into syllabic nuclei, pause detection, pitch range detection, analysis of syllable-internal pitch variation, and detection of pitch levels. The resulting system is evaluated in section 6. This is followed by a discussion of the properties of the system, its limitations and advantages, and the future perspectives for its extension. The final section gives a conclusion.

2 Pitch range and the characterization of pitch contours

When considering the choice of a prosodic transcription, it is useful to recall the *two approaches to the characterization of pitch contours* distinguished by Ladd (1996, p. 253-256, and 2008, p. 188 ss.).

The “initializing models”, on the one hand, describe pitch contours as *configurations*, i.e. as a sequence of components defined in a relational way, relative to what precedes, e.g. a contour consisting of low unstressed syllables, followed by a large fall in the stressed syllable. Such configurations are characterized by the location, the size (melodic interval) and the direction of the pitch movements. Typically, this type of representation does not refer to pitch range, because it is assumed that a given configuration can occur at various places in the pitch range without this affecting the identity or function of the contour.

The “normalizing approach”, on the other hand, factors out pitch differences related to speaker identity or to paralinguistic aspects (e.g. bored speech, emotional speech), in order to determine the *tonal space* of a speaker (Ladd, 2008, p. 193). Autosegmental models fall into this class, but in addition to the pitch normalization, they describe pitch movements in terms of sequences of pitch targets, assuming only two basic pitch levels: low and high. This illustrates that autosegmental models make decisions about the distinctive nature of pitch variations, although these decisions are not based on falsifiable procedures.

The normalizing approaches suppose a model of the *pitch range* taking into account inter-speaker differences (in particular pitch differences between voices of males, females and children) and within-speaker paralinguistic variation of pitch range. The complexity of the pitch range model stems from the fact that “the pitch range can also be modified within the speech of any one speaker” (Ladd 1996, p. 259): “For example, the voice can be raised in anger, or to express surprise, or simply to make oneself heard in a noisy room; it can be lowered if the speaker is depressed, or to express confidentiality, or simply to keep from waking the children.”

The above citation illustrates the potential ambiguity of the term *pitch range*, which indicates either the *vocal range* (*tessitura*, *global pitch span*) of a speaker, or the *register* or *local pitch span* of a given stretch of speech of that same speaker. The pitch span may be locally reduced or widened; its central value may be locally raised or lowered. In what follows, *pitch range* will be used as a synonym of vocal range, and no distinction will be made between local pitch span and vocal range.

To characterize *pitch range* Ladd (1996, p. 260-261) first distinguishes two dimensions of variation: differences of *overall level* (which may be approximated by the mean or the median f0 of the speaker) and *span* differences (the range of f0 used by the speaker). The latter may be expressed in various ways, for instance on a logarithmic scale (in semitones), to measure it independently from the overall level. Some speakers use a wide span, others a narrow one. Hirst (2011, p. 71) makes a similar distinction, but uses the term *key* to refer to the central point for the speaker's pitch range. As Ladd observes, overall level and span are often conflated because changes of level and changes of span may be hard to distinguish in some cases: raising a high tone produces the same effect as widening the span (Ladd, 2008, p. 198). Pitch range measures may be used in the normalization of pitch contours. In turn, normalized pitch contours are useful because they show a high degree of inter-speaker agreement and because they show the invariance of pitch contours pronounced by the same speaker in different paralinguistic conditions.

Ladd (1996, p. 267-269; 2008) examines several *quantitative models of pitch range*. The simple ones involve multiplicative components for range (span), bottom pitch (lower frequency of a voice) and target, whereas the more complex ones involve an additive component to shift the range upwards or downwards. He concludes more experimental data is required to formulate an adequate model of pitch range. De Looze and Hirst (2010) show that (when pitch data is stylized using Momel and characterized in terms of INTSINT targets) a strong correlation is found between the extrema of the global pitch range, on the one hand, and, on the other hand, distances from the median of the pitch targets. They also propose an algorithm to detect changes of local pitch range.

As Ladd (1996, p. 267) notes, "the *bottom of the speaking range* is a fairly constant feature in an individual's voice" (cf. Ladd, 2008, p. 203). 't Hart (1998, p. 100) states: "within one speaker, the utterance-final frequency varies very little". As can be readily observed, listeners are able to detect the bottom pitch of a speaker on the basis of phonation characteristics (Honorof and Whalen, 2005). As the pitch approaches the bottom of the pitch range, the vibration frequency moves away from the characteristic frequency of the vocal fold vibration, and as a result this vibration may become irregular, as for creaky voice, or breathy. (Of course, creak is not restricted to low-pitched phonation, but their association may be conventional in some language variants or for some speakers.)

Pitch movements may be abrupt or gradual. The first occur at the boundary between syllables, while the latter are spread over sequences of syllables (cf. 't Hart, 1998, p. 96) or appear within one and the same syllable (*glissando*). In phonological studies such pitch variations are commonly characterized in terms of pitch levels and/or pitch movements. When movements are used, their position within the syllable or relative to the vowel onset is taken into account, resulting in a distinction between early, late and very late movements (cf. ‘t Hart, 1998, p. 96).

It could be argued that the *Polytonia* transcription combines the two approaches to pitch contour characterization described by Ladd. Similar to the “normalizing” models it takes into account pitch range and it uses pitch levels to identify the special status of the bottom and top pitch, but also to partly characterize contours. At the same time, the transcription shares with the “initializing” models the relational definition of contour components: in the transcription, for instance, intra-syllabic pitch movements are characterised by the starting pitch level, their direction and size.

The next section will describe the actual annotation of tonal events which is targeted by the automatic annotation system presented later, in section 4.

3 The proposed symbolic tonal annotation

The tonal transcription identifies pitch level, pitch movement direction and size, as well as compound movements. Illustrations are given in the lower tier of figures 4 to 6, where the symbols “*L*”, “*M*”, “*H*” and “*LR*” indicate respectively a low, mid and high pitch level, and a large rise starting from a low pitch level. Before listing the full set of symbols and their possible combinations, the choices for each aspect will be motivated. In this annotation, symbols for level and movement are typically associated with each syllable, and this suggested the name *Polytonia*.

3.1 Pitch levels

As noted, the proposed annotation defines *pitch levels* in two ways: *locally*, i.e. relative to the context, and *globally*, i.e. relative to the speaker’s pitch range.

The *local* interpretation results in pitch levels low (L), mid (M) and high (H). In this case, pitch levels are defined relative to one another. For instance, when a syllable is said to be on a high pitch level, its pitch is considerably higher than some other syllable in the context and somewhere between these two syllables there has been a pitch movement, either a rapid or a gradual one, which is sufficiently large to trigger a change of pitch level. Pitch movements may occur either between adjacent syllables or within a single syllable. As a result, pitch levels are defined locally, on the basis of the sequence of pitch changes occurring between or within syllables.

The *global* interpretation provides two additional pitch levels: the top (T) and bottom (B) of the pitch range. The estimation of pitch range is described in section 5.5. These additional levels are motivated by their specialized function in spoken language. In many languages, the bottom pitch marks the end of a maximal prosodic unit or the end of a speaker turn, and the top level is mostly reserved for emotional or emphatic speech.

An important consequence of the relative successive definition of pitch levels is that two or more syllables at some pitch level, say at a high level, and located at different points in the utterance, need not have the same fundamental frequency, but may differ considerably, provided there are local pitch changes that motivate these differences. In read speech one often observes an overall trend of decreasing pitch (which is known as the *declination line*) and the local pitch changes may be viewed relative to this reconstructed reference line, in such a way that the actual pitch of high syllables decreases as the utterance moves on. This doesn't hamper the interpretation in terms of pitch levels, since it is based on local changes, not on fixed, global levels.

3.2 Pitch intervals

The *pitch interval* separating the low and high pitch level varies from one speaker to the other and is determined partly by the pitch range used by the speaker. As mentioned earlier, each individual voice may be characterized by its central pitch and its pitch span. The central pitch opposes low pitched and high pitched voices. The pitch span indicates the interval between the lower and upper pitches used by the speaker in modal speech. In a one hour corpus of 42 speakers (a subset of the "Rhapsodie" corpus for French, containing 20245 syllables, cumulated syllable time 3910 s or 65 minutes), the observed pitch range for individual speakers ranged from 5.9 to 18.9 semitones, when measured according to the procedure of section 4.5. This illustrates that pitch range varies considerably between speakers, affecting the size of the pitch intervals. This variability of pitch range calls for an interpretation of pitch intervals which is relative to the speaker's pitch range.

How many categories of pitch intervals are needed for an adequate transcription of pitch movements and pitch levels? The answer to this question will depend upon the descriptive model. For phonological models with just two pitch levels, in principle one size of pitch interval suffices. The IPO model ('t Hart et al., 1990), on the other hand, opposes full-size and half-size pitch movements. INTSINT makes a distinction between large and small pitch intervals, where the latter typically occur in "down-stepping" or "up-stepping". The RaP (Rhythm and Pitch) system (Dilley et al., 2006) opposes large and small pitch excursions (the latter are indicated by '!').

Small size pitch intervals need not be limited to down-step phenomena. In some languages, a given word sequence may be interpreted as having two distinct syntactic structures, depending on the relative size of the pitch rises occurring in that sequence. This is illustrated by the well-known example for French "Il a peint la jeune fille en noir", resulting in two different readings, either "He painted the girl dressed in black" or "He painted the girl using black paint". Examples of this type suggest that for the characterization of prosody in French, a single size of pitch interval is not sufficient. However, it is not clear how to determine the adequate number of distinct pitch intervals.

One possible criterion is the existence of minimal pairs: two sound sequences that are identical, except for the size of the pitch change at a given point in the sequence. The example for French, then, is a minimal pair suggesting the need for at least two sizes of pitch intervals.

The transcription system described here distinguishes two sizes of pitch intervals: large and small. How do we determine the actual size (in ST) of these intervals? There is little research on the perception of pitch intervals of different sizes, in continuous speech. Still, it is clear that the size of functional pitch movements should exceed that of intrinsic pitch variation (typically between open and closed vowels) and co-intrinsic variation (due to phonetic context, such as voiceless obstruents). Therefore, the size of the small pitch interval is set to 3 ST, which is slightly larger than the size of microprosodic variations (Rossi et al, 1981).

For a syllable with an audible *pitch variation* (a *glissando*), pitch level will change during the syllable. For instance, consider a large rise starting on a low pitch level and ending on a high pitch level. There are basically *two ways to represent this*: either as a sequence of pitch levels or as a combination of a pitch level and movements. The first approach indicates the sequence of pitch levels reached during the syllable (as "*LH*", in our example). The second approach indicates the initial pitch level and the successive pitch movements, if any (as "*LR*", in our example). This requires symbols representing movements: *R* (large rise), *F* (large fall), *r* (small rise), *f* (small fall) and _ (flat). The first analysis is common in autosegmental models, which indicate one or more tones for a syllable. The second approach is found in some American structuralist work (e.g. Smalley, 1964) and is reminiscent of the characterization of contours in the British school (e.g. Crystal, 1969).

The pitch level sequence approach has two major drawbacks. The first one concerns the number of pitch levels needed for a narrow representation of pitch events. Given the distinction between large and small melodic intervals, the number of distinct pitch levels required to represent sequences of such intervals would be unreasonably high. For instance, to distinguish a large high fall ("*HF*") from a small high fall ("*Hf*"), and a large high rise ("*HR*") from a small high rise ("*Hr*"), one would need additional pitch levels below and above the high level (↓*H* and ↑*H*, respectively), but distinct from the low and top pitch level. To distinguish "*HF*", "*Hf*", "*HR*", "*Hr*", "*LF*", "*Lf*", "*LR*", and "*Lr*" at least 8 levels would be required: *L*, ↑*L*, ↓*L*, *H*, ↑*H*, ↓*H*, *B*, *T*. This would lead to a multiplication of levels. The second drawback concerns the inability of the pitch level sequence approach to represent the presence and location of level plateaus. For instance, the shapes "*LR_*" and "*L_R*", in which a plateau either follows or precedes a rise, as well as the "*LR*" movement without a plateau, are all represented as the pitch level sequence "*LH*". In an autosegmental framework, this distinction requires the introduction of an additional property, such as the temporal alignment of the pitch target.

A more concise and readable representation is obtained by indicating the pitch level at the start of a syllable, followed by the successive pitch movements in that syllable, if any.

3.3 Symbols used in the annotation

The notation used here is designed to meet the following *requirements*. First, it should indicate whether a syllable presents an *audible pitch variation* or not, in other words it should specify whether its pitch is flat (level), rising or falling. Second, the notation should distinguish between *large and small movements,* both for rises and falls. Third, the notation should allow for *compound movements* such as a rise-fall or a rise following a level part, and so on. Fourth, it should indicate the pitch level of *each syllable,* relative to the range of the speaker and the context. (We return to the latter point below.)

All this information can be conveyed by a fairly small *set of symbols*. Pitch levels are represented by "*L*" (low), "*H*" (high), "*M*" (mid), "*T*" (top of range) and "*B*" (bottom of range). Pitch movements will be represented by "*R*" (large rise), "*F*" (large fall), "*r*" (small rise), "*f*" (small fall) and "_" (flat). Compound movements use a sequence of these symbols: e.g. "*RF*" (rise-fall), "*_R*" (level-rise), "*R_*" (rise-level)... There are two additional symbols with a special status. First, "*S*" (sustain) indicates a syllable with a uniform level pitch and minimal duration of 250 ms, a marked contour which is fairly rare. This symbol differs from the level part "_", which may be shorter than 250 ms and may be part of a complex pitch movement such as "*_R*". Second, the symbol "*C*" (creak) indicates a syllable with creak. These symbols may be combined as indicated in table 1.

Although the annotation is able to represent compound pitch movements of any complexity, compound movements are fairly rare, even in spontaneous speech. This will be illustrated by the frequency of pitch movements in the reference corpus (table 6). So, in practice, the tonal annotation is much lighter than suggested by table 1.

Table 1

Symbols used in the tonal annotation and their combinations. The horizontal lines above "S" and "C" indicate restrictions on the combination of these symbols, e.g. by definition "S" may not be followed by another symbol.

modifier (special cases)	start pitch level	first pitch movement	next pitch movements	end pitch level
	T (top)	_ (level, flat)	*void or*	*void or*
	H (high)	R (large rise)	*same set as*	,T (top)
	M (mid)	F (large fall)	*for first*	,B (bottom)
	L (low)	r (small rise)	*movement*	
	B (bottom)	f (small fall)		
		S (sustain)		
C (creak)				

Provided f0 is detected for a given syllable, its pitch movement is always identified, since 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 a case the pitch movement will be shown without the pitch level, for instance "*R*", "*RF*", and so on.

For conciseness a short-hand notation is used. When pitch movement is level and simple (i.e. not compound), 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. Moreover "_" (flat with missing pitch level) is skipped altogether.

In general, pitch level is indicated only for the start of the syllable (more precisely, at vowel onset). As an exception to this rule, in some cases the *pitch level reached at the end of the syllable* is indicated as well. This is the case for bottom and top, when the syllable's pitch contour starts at a pitch level other than bottom or top. This special treatment is justified by the following observation. An intra-syllabic fall ending at the bottom level acts as a terminal boundary, in the same way as a syllable with a flat contour starting at the bottom level. For contours such as "LF", "Lf" and "HF", it may be the case that they end at the bottom, but in order to make this clear, this should be marked explicitly: "LF,B", "Lf,B" and "HF,B". Similarly, a syllable which rises to the top level has the same effect as one starting at the top. (A terminal boundary ends the largest prosodic domain. Most descriptions of prosody observe that declarative utterances end at the bottom pitch level, suggesting this function of the bottom level may be common to a large number of languages.)

The above tonal transcription does not include a *symbol for stress*. (One could use the IPA stress mark which would be placed before the first tonal symbol of a syllable.) Stress is not marked because it is not detected by the system described in this study. To our knowledge, all stress detection schemes are language-dependent, whereas our goal is a language-independent annotation. Moreover, the presence of stress introduces a reference point in prosodic structure, and if stress is marked, it would be desirable to mark other points in prosodic structure too, as well as their impact on the treatment of tones at particular places of that structure.

3.4 Underlying assumptions

Any transcription makes *theoretical assumptions* about the nature of the objects being transcribed. This also holds for a tonal annotation and even for a stylization based on tonal perception. The proposed transcription implies the following assumptions.

A. The listener has the ability to distinguish pitch movements occurring between adjacent syllables from those occurring within one and the same syllable. Listening experiments with natural or synthetic speech (Rossi, 1971; Rossi et al. 1981; Mertens et al., 1997) show that many intra-syllabic pitch movements observed in speech are indeed perceived as glissandi. House (1990) shows that a given pitch movement (say a rise with a given excursion size and slope) is perceived differently, either as a movement or as an abrupt change, depending on its location relative to the vowel in the syllable.

B. The listener has the ability to distinguish large pitch changes from small ones and categorizes them relative to the pitch span. This assumption is implied by all prosodic transcriptions and models that indicate down-step (including most autosegmental models), since the pitch interval of a down-step is smaller than that of a pitch movement between a high and a low target.

C. The listener has the ability to discriminate between simple and compound pitch movements occurring within a syllable. This claim follows from the differential glissando threshold (d'Alessandro and Mertens, 1995), which is integrated in the perceptual stylization used here.

D. Pitch range limits (bottom and top) may be identified as such by the listener (Honorof & Whalen, 2005) and they have a function which is different from that of pitch contrasts *within* the pitch range, i.e. between high and low targets. For instance, the bottom pitch level marks the end of a major prosodic domain.

E. Intra-syllabic pitch movements are synchronized with the onset of the vowel and spread over the syllable rhyme. This assumption is related to the distinction between early and late pitch movements in the IPO model ('t Hart et al., 1990).

To summarize, how a pitch movement is perceived depends not only upon its *size* and *shape*, but also on its *temporal alignment* relative to the segmental layer (in particular the vowel onset and the end of the rhyme).

Demonstrating each of these assumptions in a convincing way would require substantial additional experimental research on pitch perception, something which lies outside the scope of this study.

When we compare these assumptions to those made by other transcription systems, the differences appear to be very important. For instance, INTSINT only makes assumptions B and D, but would reject assumptions A, C and E. ToBI would reject all of the above assumptions, except B.

4 Automatic labeling of prosody

Approaches for *automatic* labeling of prosody can be classified along various criteria.

First, the amount of *information used* in the process can vary considerably. Let us recall that an utterance (a sequence of words with a syntactic structure) may be pronounced in many ways, using different intonation patterns, affecting the meaning of the resulting utterance. When labeling prosody, ideally only acoustic information should be used, since the goal is to represent the particular *prosody as it is used by the speaker*, as manifested by the speech signal (Tamburini & Caini, 2005, p. 34; Kochanski et al., 2005). In practice, however, other sources of information are often included to improve the results (e.g. Wightman and Ostendorf, 1994; Ananthakrishnan and Narayanan, 2008;

Rosenberg, 2010; Jeon and Liu, 2012). These include phonetic information (phoneme and syllable alignment), morphological information (word boundaries, POS), syntactic information (syntactic function, constituents, chunk boundaries, sentence boundaries), as well as prosodic information (the location of "word stress" or lexical stress). Using sources of information other than the speech signal itself introduces a strong bias. For instance, a system using lexical information such as lexical stress position, actually predicts lexical stress, rather than "sentence stress" (actual prominence). Similarly, the use of word boundaries partially supplies what ideally should be detected by the system itself, i.e. the location of the boundaries.

Second, the *type of transcription* obtained at the output differs according to the system: it may be broad (e.g. Wightman and Ostendorf, 1994; Campione and Véronis, 2001) or rather narrow (e.g. INTSINT as in Campione et al., 2000). Ananthakrishnan and Narayanan (2008) and Jeon and Liu (2012) retain only two possibilities for pitch accent (presence/absence) and only two for boundaries (presence/absence). Clearly, such classes result in a coarse representation of prosody which does not allow for the identification of pitch contours.

Third, the *algorithms used* range from rule-based systems (Mertens, 1987a, 1987b, 1989; Campione et al., 2000; Campione and Véronis, 2001; Bartkova et al. 2012) to supervised learning techniques, including HMM (Geoffrois, 1995), decision trees (Wightman and Ostendorf, 1994), neural networks, and so on (Braunschweiler, 2005; Ananthakrishnan and Narayanan, 2008; Wagner, 2009; Rosenberg, 2010). Jeon and Liu (2012) give a comparative overview indicating techniques (algorithms) and detection accuracy. Supervised learning techniques require a validated corpus, in order to train the system. In the case of prosody, such corpora are very rare, even for the widely used ToBI annotation for English (for which most studies use the same Boston University Radio News corpus), and they are simply lacking for fine-grained prosodic annotations, such as INTSINT.

Finally, the *features* used as the input for the recognition system play an important role. Some systems include pitch stylization (for an overview of stylization, see Hermes 2006), which may be based on perceptual properties (d'Alessandro and Mertens, 1995; Mertens, 2004b) or on acoustic properties only, as in the Momel system (Hirst et al., 1991).

When comparing the results of alternative automatic annotation systems, the results will depend largely on the above mentioned factors. For instance, detecting 5 pitch levels in combination with 5 pitch movements (and even complex movements) is considerably more complex than detecting the type of pitch accent only.

The following section describes a system for automatic tonal annotation which provides a very narrow transcription of pitch events (defined in section 3), is strictly bottom-up and does not require a training corpus. These features make it very different from most other systems.

5 Description of the automatic procedure for tonal annotation

The automatic annotation of tonal features includes several processing steps, which are described in detail in the following sections. (1) Parameter extraction. (2) Segmentation of the speech signal into syllabic nuclei, based both on the corpus annotation (phonetic alignment, syllable boundaries) and the acoustic parameters (intensity, voicing). (3) Pause detection. (4) Pitch stylization based on a model of tonal perception. (5) Detection of global pitch-range and derived pitch intervals. (6) The actual detection of tonal features includes the following steps. (a) Detection of intra-syllabic pitch movements. (b) Detection of pitch range related pitch levels: bottom and top. (c) Detection of pitch levels on the basis of local changes in pitch. (d) Detection of pitch levels on the basis of intra-syllabic pitch movements. (e) Extrapolating pitch level information. (f) Treatment of pitch plateaus.

5.1 Parameter extraction

All acoustic parameters are calculated using algorithms provided by the Praat software package (Boersma & Weenink, 2012), with their default settings, except for the time step, which is set to 5 ms. Fundamental frequency (f0) is measured using the modified autocorrelation method of Praat. The V/UV decision is derived from the fundamental frequency measurement. Creak detection is only available for corpus files providing an annotation of creak. (Obviously, pitch detection errors, when present, will affect later processing stages, as is the case for all annotation systems based on f0.)

5.2 Segmentation into syllabic nuclei

The *syllable* is a central unit for the characterization of many aspects of prosody, including prominence, stress, syllable duration, speech rate, rhythm, and pitch movements. While some pitch events are synchronized with individual syllables (in particular with the stressed syllable), others are carried by a sequence of contiguous unstressed syllables (‘t Hart, 1998). Even for gradual movements extending over several syllables, the boundaries of the movement coincide with those of a syllable. Hence it is necessary to locate syllables and syllable boundaries.

The syllable is of course a phonological unit. From a phonetic and an acoustic perspective, the exact location of the *boundaries between syllables* is sometimes unclear. Whereas plosives and voiceless fricatives are characterized by abrupt changes in the acoustic signal, phoneme boundaries are acoustically less clear for glides, nasals and laterals, and for such sounds it is hard to say where exactly the boundary lies. Also, whereas many phonological theories of syllable structure assume that each sound belongs to one syllable only, from a production and perception perspective many sounds may be ambisyllabic: the closure of the consonant is part of the coda of a first syllable, whereas the release of that same consonant starts the onset of the next syllable. To avoid these issues we will not attempt to detect syllable boundaries, but rather detect the syllabic nucleus, as defined below.

A syllable may contain both voiced and unvoiced segments. Fundamental frequency is only defined for the voiced portion. Moreover intensity and spectral distribution may vary considerably during the syllable. These changes in periodicity, intensity and spectrum affect the perception of pitch changes (e.g. Rossi, 1978). So which part of the syllable is relevant for the calculation of its prosodic properties?

The start of the vowel constitutes an anchor for the *temporal alignment of pitch events* with respect to the sequence of speech sounds. The IPO model of intonation (‘t Hart et al., 1990), for instance, distinguishes between early and late pitch movements, and this temporal alignment is expressed relative to the start of the vowel. As mentioned earlier, House (1990) shows a pitch variation is perceived either as a movement or as an abrupt step, depending on its location relative to the vowel. On the other hand, it can also be shown that when a particular pitch movement, such as a rise or a fall, is associated with a given syllable, the actual shape of this movement will depend to some extent on the composition of the syllable, more precisely on the number and nature of the consonants in the coda.

For acoustic analysis, one has to decide which part of the syllable is used for calculating the prosodic properties. The unit which is used here may be broadly characterized as the central part of the voiced area of a syllable, located around its local peak of intensity, for which the intensity only decreases to some amount, specified by a threshold. This central part of the syllable will be referred to as the *syllabic nucleus*. (Note that this term is used differently in phonology.) Given the above definition based on acoustic parameters, the syllabic nucleus may be used for the automatic segmentation of the speech signal.

The syllabic nucleus has been used in the context of automatic segmentation for the analysis of prosody in Mertens (1987b, 1989, 2004a, 2004b). Tamburini and Caini (2005, p. 35) use a similar notion; they show that syllable duration and syllabic nucleus duration have similar distributions and suggest that the latter can be used as a replacement for syllable duration in the measurement of syllable lengthening.

For the purpose of evaluation the definition of the syllabic nucleus is further constrained. When evaluating the tonal annotation system, one wants to avoid errors due to syllable segmentation in order to evaluate the tonal annotation independently of the segmentation. For this reason, the segmentation into syllabic nuclei will be guided by the phonetic and syllabic alignment provided by the corpus annotation. The latter will be used to locate the part of the syllable where the syllabic nucleus is to be found, more precisely to locate the time interval of the *syllable rhyme*, consisting of the vowel followed by the coda.

Thus, in the present system the syllabic nucleus is located as the local peak of intensity within the voiced portion of the rhyme of the syllable as provided by the corpus annotation. The left and right boundaries are identified using a threshold for intensity drop. For the left side, this point is at 2 dB below the peak. For most recording conditions this is an acceptable approximation of the start of the vowel. For the right side, the intensity drop is equal to 80% of the intensity difference between the peak and the local minimum between the peak and the end of the rhyme, with a minimum of 3 dB. This latter threshold is chosen in such a way as to include within the syllabic nucleus the voiced part of the rhyme, for which intensity is relatively high and which is perceived with a clear pitch. Also, it allows for correcting errors in the phonetic alignment provided with the corpus.

When the resulting time interval contains a rapid pitch change, most often resulting from an octave jump (pitch detection error), the interval will be truncated at this discontinuity. Moreover, a minimal duration for the resulting syllabic nucleus is required, otherwise it is rejected. The resulting segmentation is illustrated in figure 1.

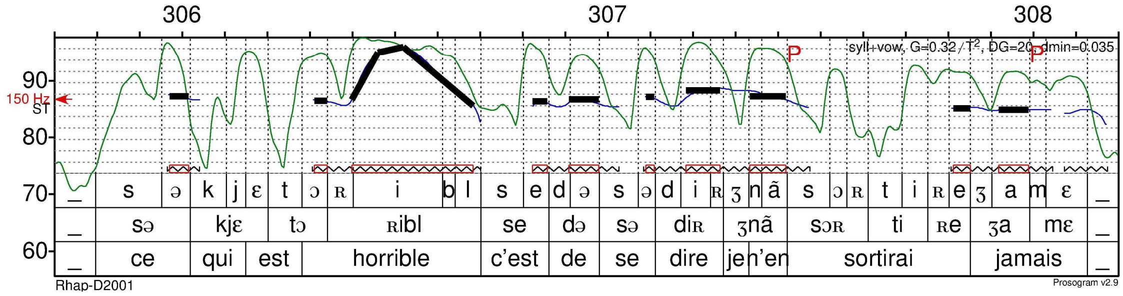

Fig. 1. Illustration of the segmentation into syllabic nuclei guided by the phonetic alignment, for the utterance “ce qui est horrible c’est de se dire je n’en sortirai jamais” (“What is horrible is to say to oneself: I will never get out of this situation”). The upper part shows the acoustic parameters of intensity (continuous thin green line), voicing (saw tooth), fundamental frequency (thin blue line, mostly covered by the thick black line), as well the pitch stylization (thick black line) described in section 5.4. Pitch is plotted on a semitone (ST) scale (relative to 1Hz), with horizontal calibration lines at 2ST steps. The lower part shows various corpus annotation tiers: phonetic alignment, syllables and words. The syllabic nuclei appear as red boxes on top of the voicing line (saw tooth). For the syllable [ʀibl], the nucleus includes the vowel and part of the coda. The syllables [kjɛ], [sɔʀ] and [ti] are analyzed as unvoiced and consequently no nucleus is detected.

5.3 Detection of pauses

Silent pauses play an important role in discourse, for instance in the marking of boundaries between syntactic constituents or in speaker turn management. Silent pauses also affect the perception of pitch events, by lowering the glissando threshold (House, 1995, 1996).

In order to take into account this effect, speech pause detection is needed. Although pause detection may seem straightforward, it is often complicated by background noise and speech dynamics which depend on recording conditions and which vary from one corpus to the next. Most algorithms for pause detection are based on the intensity difference between the (average) level of the voiced speech intervals and the level of the silent pauses. In the case of background noise, the level of the non-speech interval increases. In the case changing recording conditions (e.g. gain adjustments during the recording), it is not possible to select an optimal threshold.

As a practical solution, the following procedure is used. The segmentation into syllables (cf. supra) results in a sequence of syllabic nuclei. The gap between the end of a nucleus and the beginning of the next may be used as a rough estimate of the pause length. When this gap exceeds 350 ms, it is interpreted as a pause. In the figures detected pauses are indicated by a "P" placed at the start of the gap.

5.4 Pitch stylization

The next step applies a stylization to the f0 data (cf. section 5.1) of each of the syllabic nuclei obtained by the segmentation (cf. section 5.2). This pitch stylization is based on a model of *tonal perception* in speech. It has been described in great detail elsewhere (d'Alessandro and Mertens, 1995; Mertens, 2004a, 2004b) and has been validated in listening experiments using resynthesized speech (Mertens et al., 1997) for French. The algorithm may be summarized as follows: for each syllabic nucleus, the pitch contour is divided into one or more parts of uniform slope ("tonal segments"), on the basis of a perceptual threshold for slope change (the differential glissando threshold); for each part the pitch change is compared to the glissando threshold in order to determine whether the measured variation is perceived as a glissando or not. This results in a representation of the audible pitch events in an utterance, which is less complex than the acoustic data itself. The obtained stylization is shown in figure 1 and in later figures in this article.

In the current system, the model has been improved by taking into account the effect of the following pause (cf. section 5.3) on the glissando threshold: the presence of pause lowers the threshold.

The pitch stylization based on perceptual properties, which is used here, applies perceptual thresholds to the pitch variations within the syllabic nuclei provided by the segmentation. This approach is rather *different from* the two types of *stylization used in the IPO model* ('t Hart et al., 1990). The first type, the "close-copy" stylization, involves interactive comparisons between two resynthesized speech signals: one with the original pitch contour and a modified version, in which the pitch contour is obtained by connecting by straight lines the pitch targets supplied interactively by a phonetician. The second type of stylization, which uses "standardized pitch movements", involves an approximation of the observed pitch contour by a concatenation of shapes taken from a small inventory of 10 movements and 2 level segments (the low and high declination lines), which are assumed to provide a characterization of the pitch contour retaining all functional (distinctive) information. In contrast with the IPO approach, the procedure described here does not require manual intervention, subjective perceptual judgments, or predefined shapes.

5.5 Automatic detection of the pitch range of a speaker

Information about pitch range will be used in several ways. First, it is used to detect and discard pitch values outside the pitch range of the speaker. Second, it is used to assign a pitch level to pitch values near both ends of the global pitch range. Finally, it is used to adjust the thresholds for melodic interval categories (small versus large intervals) to the pitch span used by the speaker.

The *estimation* of the pitch range (both span and central pitch) is based on all the syllables pronounced by a speaker. However, potentially unreliable values are discarded to improve robustness. This is done in two steps. First, syllables containing octave jumps (detected as discontinuities), syllables pronounced with a creaky voice as well as those labeled as hesitations (in the corpus annotation) are discarded from the data used for pitch range calculation. The second step aims at removing pitch values which are unreliable, because too far away from the center of the pitch range, i.e. from the median of the pitch data for the speaker. Syllables with pitch values that differ from the median pitch by at least 18 ST are discarded. In uncontrolled speech, pitch range most often is smaller than 18 ST.

For the retained syllables a simple statistical measure is applied. For each syllable pronounced by the speaker (as identified from the annotation), two pitch values are obtained: the minimum and maximum pitch inside the syllabic nucleus. This results in the distribution of pitch values for a speaker. The 2% and 98% percentiles of these measured values provide an estimate of the bottom and the top of the global pitch range, respectively. By using these percentiles, rather than all the observed minimum and maximum values of the data, extreme values which may be due to pitch detection errors are mostly eliminated. Limiting pitch values to those in the syllabic nucleus eliminates most pitch measurement errors related to co-intrinsic pitch phenomena.

The *validity* and robustness of this estimate of a speaker's pitch range largely depends upon the number of syllables pronounced by that speaker in the corpus. In order to obtain representative values, some 200 to 300 syllables are needed. Therefore the pitch span estimation procedure should not be used for isolated utterances.

The global pitch range differs considerably among speakers: in a 116 min. corpus of French in various speech types containing 42 speakers, the measured pitch span for individual speakers ranges from 5.9 to 18.9 semitones.

A closer look at the distribution of the median pitch value (cf. Ladd's *overall level*, Hirst's *key*) and the pitch range for various speakers shows an interesting observation, also illustrated by figure 2: the position of the median pitch is not necessarily located near the center of the pitch span (on a scale in ST), but varies from one speaker to the other, possibly resulting in a skewed pitch range, where distance below the median is much smaller than that above it (negative skew), or vice versa (positive skew). For practical reasons, we will use the following terms. The *lower part* of the pitch span, below the median pitch, may be roughly divided into the *first and second quarter* at a point halfway between the bottom and the median. Similarly, the *upper part* will be divided into the *third and fourth quarter* at the point half-way between the median and the top.

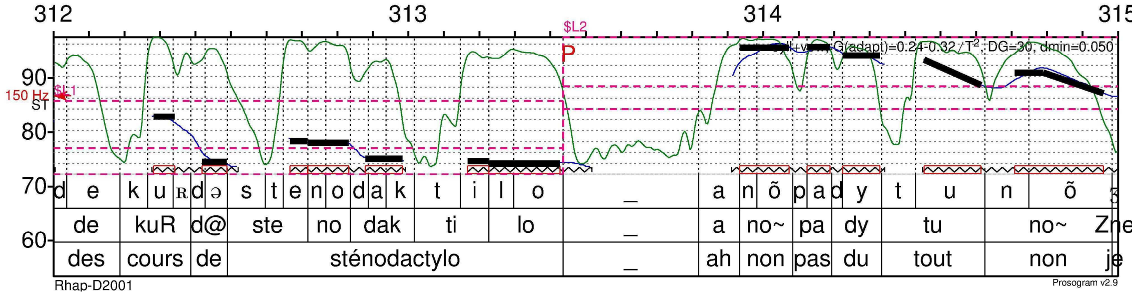

Fig. 2. Illustration of pitch range detection, for the utterances “des cours de sténodactylo – ah non pas du tout non” (“courses of shorthand typing – ah no, not at all, no”). Acoustic parameters and corpus annotation appear as in figure 1. The detected pitch range for a given speaker is represented by three horizontal dashed lines in red, indicating the estimated top and bottom of the range as well as the median of all pitch values by the speaker (line between top and bottom). At the speaker turn near 313.4s, the detected pitch range changes from a low pitched voice (male speaker, L1) to a high pitched voice (female speaker, L2).

Pitch intervals are used differently by speakers depending on the size of their pitch span. Obviously a speaker with a narrow pitch span will use smaller pitch intervals, in order to maintain the opposition between *small and large intervals* while using a narrow range. The estimated pitch span is used in the system to define categories of melodic intervals and to categorize pitch movements and pitch levels in later steps of the tonal annotation procedure. The pitch interval sizes were determined empirically on the basis of the analysis by the author of a corpus containing 42 speakers with different pitch ranges. The following table shows the minimum pitch change for large and small intervals, depending upon the speaker’s pitch range.

Table 2

Thresholds for large and small pitch intervals used for pitch movement classification and pitch level determination, depending on the pitch range obtained for a given speaker, following the procedure described in the text.

Pitch range	Large interval	Small interval
> 8.5 ST	> 4.5 ST	3.0 - 4.5 ST
7.0 – 8.5 ST	> 3.5 ST	2.5 - 3.5 ST
< 7.0 ST	> 3.2 ST	2.5 - 3.2 ST

5.6 Syllable-internal pitch movements

The detection of pitch levels and pitch movements consists of several steps, the first of which is described in this section and deals with intra-syllabic pitch variations.

In a first stage the pitch contour is stylized as described in section 5.4. For each syllabic nucleus (cf. section 5.2), the measured pitch variation is divided into "tonal segments" with a uniform slope (either level, rising, or falling). For each tonal segment the observed pitch slope is compared with the glissando threshold. Subliminal variations (i.e. below the glissando threshold; inaudible variation of f0) are normalised to level pitch segments.

In the stylization used here, the glissando threshold is lowered for syllables followed by a pause (detected as described in section 5.3), in agreement with the findings of House (1995). For syllables followed by a pause, a glissando threshold of $0.16/T^2$ is used, corresponding to the threshold observed for isolated stimuli in psychoacoustic experiments. For syllables not followed by a pause, the glissando threshold is raised to $0.32/T^2$.

In a second stage pitch segments with an audible (supraliminal) pitch variation are further categorized in terms of large and small pitch intervals. The latter are not defined absolutely, but rather relative to the pitch range of the speaker (cf. supra, section 5.5).

As a result, we obtain the elementary forms for intra-syllabic pitch movements used in the transcription convention described in section 3: _ (level, subliminal), *R* (large rise), *F* (large fall), *r* (small rise), *f* (small fall). Compound movements are represented by sequences of elementary shapes: *RF* (rise-fall), *_R* (level then rise), *R_* (rise then level), *Rr* (large rise followed by small rise), and so on.

As mentioned in section 3 an additional elementary form *S* indicates syllables with a sustained and uniform level pitch. In this case, the syllabic nucleus has a duration of at least 250 ms, and its pitch contour is level or slightly falling throughout the nucleus, with an intra-syllabic pitch movement between 0 and -1.5 ST.

5.7 Pitch level detection

To detect the pitch level of a syllable or a sequence of syllables, various types of information are used: the speaker's pitch range, the pitch changes between successive syllables, and the intra-syllabic pitch movements. In addition, when the pitch level cannot be determined directly using this information, it may often be derived indirectly from the identified pitch level of neighboring syllables. Each of these aspects will be described in detail below.

In order to combine the various sources of information, they are analyzed one by one in the order indicated below, until pitch level is detected or until all sources of information have been used. As soon as pitch level is identified by some step in the procedure, the following steps are skipped.

For some syllables pitch level remains unidentified even after applying all steps. In such cases the information is insufficient and this is indicated in the annotation by the lack of pitch level.

5.7.1 Pitch level detection based on pitch span

For syllables where f0 starts above the estimated top or below the estimated bottom of the speaker's pitch span, the pitch level will be set to T (top) or B (bottom) respectively.

However the top level is assigned only when this can be done reliably, i.e. when sufficient (at least 200) syllables are available for the speaker in the corpus and the pitch span is sufficiently wide, more precisely when the upper part of the pitch span, between the median and the top, exceeds 8 ST. Otherwise the syllable is labeled as H (high).

5.7.2 Pitch level detection based on local pitch change

The pitch variation in the local left context of the *target* syllable, i.e. the syllable to be labeled, may be used to *infer its pitch level*. For instance, when the start pitch of the target syllable is sufficiently higher than the lower pitch value in the left context, the target is high within that context. Similarly, when the start pitch is sufficiently lower than the higher pitch value in the context, the target is relatively low. A similar rule may be made for small pitch intervals, where a mid pitch level is inferred. The pitch values used in the procedure are supplied by the pitch stylization.

The optimal *size of the left context* is selected as follows. The local context consists of up to 3 syllables preceding the target syllable and occurring within a time window of 500ms. There should be no pause between the target syllable and the syllables in the left context; the presence of a pause introduces a left boundary for the context. The context and target syllables should all be pronounced by the same speaker. Syllables tagged as hesitations in the annotation are discarded from the context: such hesitations are often pronounced on a low or bottom level. Moreover, syllables with a top or bottom pitch level are excluded from the context: the low, mid and high levels are defined relative to each other, but not relative to the bottom ot the top.

To account for all possible *pitch configurations* in the context, we calculate the melodic interval I_{up} between f_{min} , the minimum pitch value in the context, and f_{start} , the pitch at the start of the target syllable, as well as the interval I_{down} between f_{max} , the maximum pitch value in the context, and f_{start}. The target is assigned a H (high) pitch level, when I_{up} exceeds the large upward (positive) interval (as defined above, in section 5.5, table 2); it is assigned a M (mid) pitch level, when I_{up} exceeds the small upward interval (as defined above); it is assigned a L (low) pitch level, when I_{down} exceeds the large downward (negative) interval. In all other cases, no pitch level is assigned to the target syllable. The general idea is illustrated in figure 3.

For syllables in the context with a pitch movement, the procedure is the same: the minimum and maximum pitch values are used in the construction of the local pitch references. The target syllable itself may also be a pitch movement, since we are establishing the pitch at the start of the target, and the movement itself will be represented as such and not as a sequence of pitch levels.

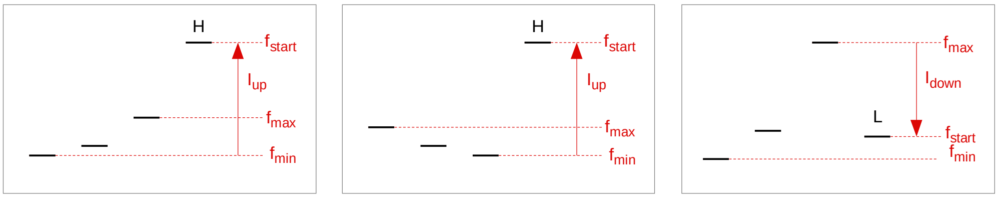

Fig. 3. Illustration of the procedure for pitch level assignment based on pitch information in the local left context. The horizontal and vertical axes represent time and pitch respectively. Each black mark corresponds to a syllable. In the examples the first three syllables make up the local left context for the fourth syllable which is the target syllable. f_{min} and f_{max} correspond to the pitch minimum and maximum in the context. I_{up} is the interval between f_{min} and the target f_{start} . I_{down} is the interval between f_{max} and the target.

5.7.3 Pitch level inferred from intra-syllabic pitch movements

When a syllable contains a large pitch variation, either simple or compound, this variation also provides information about the pitch level at the start of that syllable. This information is less certain than the local context information (cf. previous section), because it is based on a single syllable. For robustness, the information about pitch variation should be completed by information about pitch range. The following rules are used, where the *center* corresponds to the median of the pitch range. Speakers with a large pitch range may show large pitch movements (> 4.5 ST) below and above the center.

- A large rise (*R*) starting below the center is labeled as a low pitch level (*L*).

- A large rise (*R*) starting slightly above the center receives a mid pitch level (*M*).

- A large fall (*F*) starting below the center and ending at the bottom of the pitch range receives a low pitch level (*L*).

- A large fall (*F*) starting in the upper quarter of the pitch range receives a high pitch level (*H*).

- A small rise (*r*) which starts slightly higher (within a “small” interval) than the preceding syllable, which is labeled low, will be labeled as starting at a low pitch level (*L*).

- A syllable with a level pitch (_), starting in the upper quarter of the pitch range is labeled high (*H*).

As said earlier, these rules are applied only to syllables for which pitch level could not be attributed on the basis of criteria used in previous steps.

5.7.4. Extrapolating pitch level information

For some syllables pitch level remains unknown after the previous steps. In this case, the detected pitch levels in the immediate context will be used as an additional source of information. This is done by measuring the pitch interval between a target syllable with unknown pitch level (but known f0) and an adjacent or near syllable with detected pitch level, and which acts as the reference level. If this interval is very small (≤1.2 ST), the target receives the same pitch level as the reference syllable. If it is a small or a large interval (as defined above), the target syllable is labeled high, mid or low, depending on the size and direction (sign) of the pitch interval between the target and the reference. This procedure can be applied in both directions, forward and backward, and for reference syllables that are either adjacent or further away.

The procedure is applied with increasing context size, looking first for an adjacent reference, then for a more distant one, but within a time window of 0.5s separating the target from the reference. The search is interrupted as soon as the pitch level of the target is found. The procedure is applied first backward (the target precedes the reference), then forward. Obviously, the procedure is applied only to syllables pronounced by the same speaker.

5.7.5 Pitch levels for plateaus

The detection of pitch level described above relies mainly on pitch *changes*, either between adjacent syllables or within a single syllable. As a result it is not effective for *pitch plateaus*, i.e. sequences of level syllables pronounced at the same pitch level. For the latter a specialized treatment is used.

In order to locate such plateaus, the speech chain is scanned for sequences of 3 syllables for which pitch level has not been assigned and where the pitch interval between successive syllables is negligible (≤1.2 ST). In these sequences the first syllable will receive a pitch level according to its position in the pitch range of the speaker: the level is set to low, when the pitch is located in the lower part of the pitch range, and to mid, when it is located in the third quarter of the pitch range. This procedure is followed by an extrapolation step, as described in the previous section, to extend the pitch level to adjacent syllables, where possible.

5.8 The resulting tonal annotation

Figure 4 illustrates the results obtained by the automatic tonal annotation, for the French utterance “ce qui vous a toujours intéressée chez les gens c’était le mécanisme de la carrière” (stressed syllables are underlined) pronounced by a male speaker. Various aspects of the tonal annotation are shown: (1) the distinction between gradual pitch movements (such as the gradual fall on “a toujours intéressée”) and abrupt pitch changes (such as between the last two syllables of “intéressée”), (2) the distinction between syllables with a glissando (such as the rise on the last syllable of “carrière” or the fall at the end of “était”) and those with a steady pitch (such as the last syllable of “intéressée”), (3) the distinction between large and small intervals for glissandos (compare the rises on “gens” and “carrière”), (4) the distinction between the pitch level at the starting point of a pitch movement (compare the high rise on “gens” and the low rise on “carrière”). The example also illustrates the local interpretation of pitch level: syllables with a same pitch level may be at different frequencies, as is the case for he syllables labelled H in “a”, “intéressée”, “gens”, and “mécanisme”; the pitch level is determined locally, on the basis of pitch changes in the near context. Finally it shows that in gradual pitch movements, extending over a sequence of syllables, the pitch level changes from *H* to *M* to *L*, and for some syllables along this continuum the pitch level may remain unassigned.

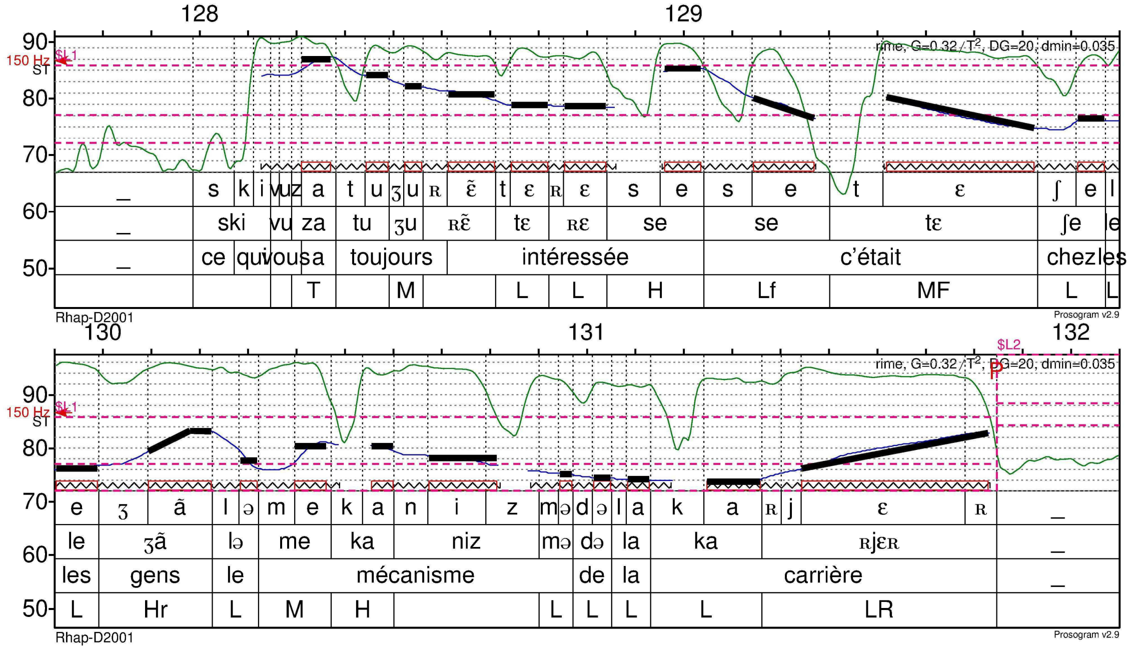

Fig. 4. Automatic prosodic labeling of the utterance “ce qui vous a toujours intéressé chez les gens c’était le mécanisme de la carrière” (“What has always interested you in people, was the career mechanism.”), by a male speaker. The automatically obtained labeling is shown in the lower tier. Acoustic parameters, corpus annotation and pitch range are shown in the same way as in earlier figures. The excerpt is commented on in the text.

Figures 5 and 6 illustrate the distinction between pitch movements with or without a level part (or plateau), opposing for instance the rise on the syllable “ans” and the level-rise-level on the last syllable of “Angelina”. This level part is represented by an underscore in the tonal annotation.

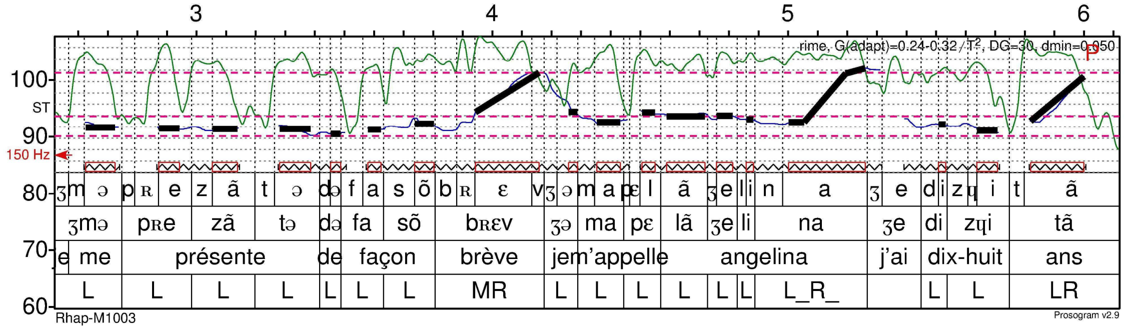

Fig. 5. Automatic tonal annotation for the French utterances “Je me présente de façon brève. Je m’appelle Angelina. J’ai dix-huit ans.” (“I briefly introduce myself. My name is Angelina. I’m 18 years old.”), by a female speaker. The automatic prosodic labeling is shown in the lower tier. Acoustic parameters and corpus annotation are shown in the same way as in previous figures.

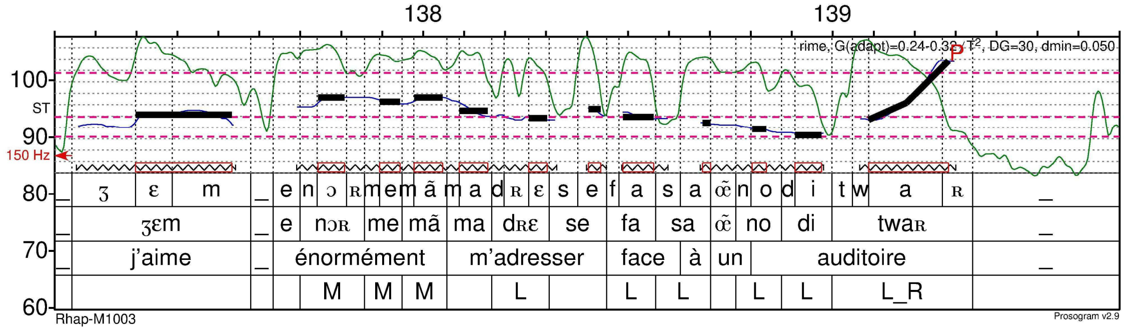

Fig. 6. Automatic tonal annotation for the French utterance “J’aime énormément m’adresser face à un auditoire.” (“I really enjoy talking in front of an audience.”) The prosodic labeling is shown in the lower tier. Acoustic parameters and corpus annotation are shown in the same way as in previous figures.

5.9 Limitations of the approach

Phonological models of intonation do not assign a tone or pitch level to every syllable in the utterance, but only to those located at *certain positions in prosodic structure*. Depending on the model, such locations include the stressed syllable, the pitch accent, the syllable at a prosodic boundary, the syllable preceding stress, and so on. Unfortunately, taking into account prosodic structure requires the prior detection of stress or prominence, and of prosodic boundaries, all of which are language-specific properties. In contrast, the tonal annotation described here avoids language-specific properties as much as possible, such that it may be applied to many languages. For this reason pitch level and movement are assigned to each syllable, even if this introduces redundancy.

When pitch level is assigned to every syllable, using a fixed number of levels, *gradual pitch movements* cannot be represented appropriately. Such gradual movements typically occur on sequences of unstressed syllables. For instance, for a gradual rise spread over several syllables, starting from a low pitch and ending in a high pitch level, there will inevitably be a syllable where the symbol indicating pitch level changes from *L* to *M* or *H*, even though the pitch interval between this syllable and the preceding one may be very small. This is as expected: the pitch level assignment is based on a threshold for pitch distance between the current syllable and the pitch in the left context.

So, although marking pitch level for every syllable is not optimal, in particular for gradual movements, it is inevitable when there is no information (such as stress) to select particular positions in the syllable chain. Still, when information about stress, prominence and boundaries becomes available in a later stage, it can be nicely integrated with pitch level information, since the latter will only be preserved for the relevant positions in prosodic structure.

6 Evaluation of the annotation system

Automatic prosodic labeling involves two major aspects, which should be separated during evaluation. The first concerns the way in which prosody is represented, i.e. the *annotation convention* or *labeling scheme*, and how this representation may be used by human labelers. The second aspect is the *automatic annotation system*. A labeling convention may be defined independently of an automatic annotation system using this convention. Both aspects should be evaluated separately.

This section only deals with the evaluation of the annotation system. However, we return to the questions about the annotation convention in section 7.

The central question here is whether the system generates the expected sequence of symbols, irrespective of the nature of these symbols (which could be linguistically motivated or not, for instance). To evaluate this, the amount of agreement between the predicted annotation and a reference annotation is measured, resulting in a global evaluation of the system.

In addition, given that the system includes several processing steps, each of these steps may be evaluated separately. Some components lie outside the scope of this study. For instance, pitch determination (section 5.1) is discussed in great detail in Hess (1983) and continues to be a research topic by itself. The phonetic alignment and syllable segmentation (section 5.2) are provided by the speech corpus; they are inputs to prosodic labeling system, not an output that should be evaluated here. Pause detection (section 5.3) relies directly on the syllable segmentation, and hence does not require evaluation. The pitch stylization algorithm (section 5.4) has already been evaluated in detail elsewhere (see references in section 5.4). Finally, the estimation of global pitch range (section 5.5) is straightforward: it uses basic statistical measures of the long term distribution of the pitch targets computed by the pitch stylization.

For these reasons, this section will be restricted to a *global* evaluation of the computational system for automatic prosodic labeling. A reference corpus will be constructed for this purpose. It is used only to evaluate to what extent the system generates the expected prosodic annotation of an utterance. Hence, the evaluation makes no claims about the adequacy of the annotation convention itself.

The evaluation of automatic annotation systems is usually based on a comparison between the annotation obtained by the system, and a *reference annotation*: the quality of the system is proportional to the agreement between both annotations. In the case of a prosodic labeling, this standard approach is hampered by the fact that, for most languages, corpora with reference annotations for prosody are hardly available. Even for languages for which prosodically annotated corpora are available, there are none with the labeling scheme proposed here (section 3). This raises the question as to how the system should be evaluated in the absence of a reference corpus.

The seemingly straightforward answer would be to provide such a reference prosodic annotation for an existing speech corpus. However, this approach has some drawbacks. The manual annotation of the corpus by one or more prosody experts using the new prosodic annotation convention would require a huge amount of time, even for a small corpus. Moreover, as with other types of annotation, conflicts between the judges are inevitable and a criterion has to be found to resolve them in order to arrive at the final reference annotation.

As an alternative, and in order to provide a reference corpus within reasonable time, the reference corpus may be obtained *semi-automatically*, by hand-correcting the output of the automatic annotation system. Obviously, this approach introduces an important bias, since the semi-automatic reference annotation will differ considerably from one created from scratch, without the input from an automatic labeling system. Still, this procedure is merely seen as a way to detect and quantify errors in the automatic annotation. We do not claim that an untrained listener would be able to provide a transcription matching the detail of the one provided by the automatic system.

6.1 Construction of the reference corpus

The reference corpus, indicating pitch level and pitch movement as described in section 3, was prepared, by hand-correcting the output of the automatic tonal annotation system.

6.1.1 Speech material

The speech material includes recordings from two languages with a different prosodic structure. French is a fixed stress language: lexical stress is on the last syllable of the word containing a vowel other than schwa. Dutch is a free stress language: the position of lexical stress varies from one word (lexical item) to the next.

Three speech recordings were selected to meet the following requirements: spontaneous speech, inclusion of male and female speakers, availability of a phonetic and syllabic alignment, relatively short duration, acceptable recording conditions, little background noise, few pitch detection problems, few cases of creaky voice, whispered or unvoiced speech. These sub-corpora will be referred to as C1, C2 and C3. The speakers of C1 and C2 are native speakers of French, those of C3 are native speakers of Dutch (Flemish variant). The French speech material was taken from the Rhapsodie corpus of spoken French (http://www.projet-rhapsodie.fr), more specifically files M1003 (=C1) and M2003 (=C2). C3 was taken from the CGN corpus (Corpus Gesproken Nederlands, http://lands.let.ru.nl/cgn/), file FV600729. The phonetic alignment was obtained semi-automatically, but validated manually by the author, following the procedure described below.

Table 3 shows some general properties of these sub-corpora. The speech rate of the speakers is fairly high. It is expressed as the number of syllables per phonation time, where phonation time indicates the duration of the speech signal (of a given speaker) without silent pauses (measured as internucleus intervals with a duration exceeding 300 ms).

Table 3

Properties of three reference corpora used for the evaluation of the tonal annotation system. Speech rate is measured in syllables/phonation time, where phonation time corresponds to the time of the speech signal without pauses. Pitch range measurements are obtained using the procedure described in section 5.4.

corpus	*lang*	*total time (s)*	*discourse type*	*speaker gender*	*speech rate (syll/s)*	*pitch range (ST)*	*bottom (Hz)*	*mean (Hz)*	*median (Hz)*	*top (Hz)*
C1	Fr	247	job interview	female	5.54	11.1	182	232	223	346
C2	Fr	294	sermon	male	5.11	16.5	105	177	172	273
C3	Du	293	radio interview	1=female,	6.03	16.3	107	162	154	275
				2=male,	6.14	12.5	78	103	97	161
				3=male	5.84	9.2	74	93	90	126

Several syllables were eliminated from the reference corpus because their f0 data was considered unreliable. These include syllables pronounced with creaky voice, those marked as hesitations, those without pitch data, and finally those pronounced during speaker overlaps (in corpus C3 only). The absence of f0 may be explained by unvoiced speech, breathy voice, vocal fry, f0 discontinuities, or too short duration of the syllabic nucleus. Table 4 indicates the numbers of syllables for the reference corpora.

Table 4

Number of syllables in the reference corpora: total number of syllables in the corpus, number of syllables with unreliable pitch values (because of creak or hesitation), number of syllables used in the evaluation of pitch level and pitch movement, and number of syllables for which a pitch level was assigned by the system.

	number of syllables			
corpus	total	unsafe	evaluated	pitch level assigned
C1	1145	259	886	693
C2	1011	16	995	788
C3	1317	134	1183	875
C1+C2+C3	3473	409	3064	2356

6.1.2 Validation procedure

The reference corpus was validated manually by one phonetician, the author of this paper, who is a native speaker of Dutch and a near-native speaker of French, with 30 years of experience in auditory transcription and acoustic analysis of speech prosody. During the validation process, he listened to short stretches of the speech signal as many times as needed to check the audibility, direction and size of the pitch movements provided by the automatic transcription, and to check whether the transcribed pitch levels corresponded to the perceived ones. To facilitate the task, various sources of information were combined, as illustrated by figures 4, 5 and 6: the speech signal itself, the phonetic alignment, the syllabic nucleus boundaries, the f0 measurement (plotted on a ST scale), and the speaker's pitch range.

6.2 Evaluation results

Except for special cases such as creaky voice, the tone label assigned by the system consists of two parts: pitch level and pitch movement. They are evaluated separately, due to the large number of combinations.

6.2.1 Evaluation for pitch level

When the signal contains insufficient information for pitch level assignment, no pitch level is assigned. For gradual uniform pitch movements (rises, falls) occurring over stretches of several syllables, this behaviour is indeed optimal, since during the change from one pitch level to the next the pitch reaches values in between two successive pitch levels, for which a decision of pitch level would be arbitrary. In other cases, the presence of a pitch level would have been preferred. In the current system, in which position in prosodic structure is not identified, it is impossible to say whether the absence of pitch level is correct (preferred) or not. Consequently, syllables without pitch level have been disregarded in the evaluation of pitch level assignment. Pitch level was assigned to 76.89% of the 3064 syllables with safe f0 data.

In order to evaluate pitch level assignment, *standard measures for classification tasks* are used: precision, recall, accuracy and F-measure. For the current task, the classes correspond to pitch levels (indicated by their labels: *B, L, M, H, T*). For each syllable in the corpus, we compare the label *predicted* by the system and the *actual* label, indicated in the reference corpus. When some label X is predicted by the system, it may correspond to X in the reference corpus (this case is called a “true positive”) or to some other label (“false positive”). The *precision* for label X indicates the fraction of all predicted labels X (true positives and false positives) that are labelled X in the reference corpus (true positives). The *recall* for label X indicates the fraction of all syllables labelled X (true positives and false negatives) in the reference corpus that are predicted as X by the classification system. The *accuracy* indicates the fraction of predicted labels in the corpus that are either true positives or true negatives. The *F-measure* combines precision and recall: F = 2•(Prec•Rec)/(Prec+Rec).

The values obtained for these measures, as well as the number of syllables in each class (i.e. each pitch level) in the reference corpus, are shown in table 5, for each sub-corpus and for the combined corpora. As can be seen from the number of syllables per class (row labeled “N”), syllables are distributed unevenly over the various pitch levels in the reference corpora. Out of a total number of 2356 syllables, 1280 syllables carry level *L* and only 13 syllable level *T*. Although this is as expected (the top of the pitch range is hardly used is modal speech), it results in sparsity problems and in unreliable values for classes with few instances, such as *T* and *B*. However, for classes *L*, *M* and *H*, with sufficient instances, the results obtained for the various measures are very high.

Table 5

Precision, recall, accuracy and F-measure for the pitch level classification obtained for reference corpora C1, C2, and C3, and for the combined data. N indicates the number of syllables in the reference corpus that belong to a given class.

C1:

	B	L	M	H	T
N	3	410	221	59	0
Prec	0.176	0.987	0.873	0.617	
Rec	1.000	0.902	0.869	0.847	
Acc	0.980	0.935	0.918	0.942	
F	0.300	0.943	0.871	0.714	

C2:

	B	L	M	H	T
N	14	379	205	189	1
Prec	0.900	0.924	0.795	0.773	0.000
Rec	0.643	0.931	0.717	0.862	0.000
Acc	0.992	0.930	0.878	0.906	0.999
F	0.750	0.928	0.754	0.815	0.000

C3:

	B	L	M	H	T
N	13	491	243	116	12
Prec	0.722	0.975	0.947	0.799	1.000
Rec	1.000	0.949	0.889	0.957	1.000
Acc	0.994	0.958	0.955	0.962	1.000
F	0.839	0.962	0.917	0.871	1.000

C1-3:

	B	L	M	H	T
N	30	1280	669	364	13
Prec	0.556	0.963	0.877	0.752	1.000
Rec	0.833	0.929	0.830	0.890	0.923
Acc	0.989	0.942	0.919	0.938	1.000
F	0.667	0.946	0.853	0.815	0.960

6.2.2 Evaluation for pitch movement

A total number of 3064 syllables were used in the evaluation of pitch movement detection. As can be seen from table 6, intra-syllabic pitch movements are distributed very unevenly in the corpus. Level tones occur in 91.16% of the syllables. Compound tones represent 0.97% of the syllables. Major pitch movements (either simple or compound) are found on 4.63% of the syllables in the corpus. The classification results are very good, especially for level tones and major pitch movements. They are slightly less for small interval movements.

Table 6

Values for 4 classification measures (precision, recall, accuracy, F-measure) computed on the automatic labelling of intra-syllabic pitch movements occurring in the combined reference corpus (C1, C2 and C3, 3064 syllables). The columns indicate the pitch movements according to the notation described in the text (where '_' indicates a level part). N indicates the number of syllables in the reference corpus for the corresponding pitch movement. % indicates the proportion of syllables with the given pitch movement.

	_	r	R	_R	R_	f	_f	F	_F	F_	_F_	S
N	2793	47	63	4	2	62	9	58	12	2	1	11
%	91.16	1.53	2.06	0.13	0.07	2.02	0.29	1.89	0.39	0.07	0.03	0.36
Prec	0.992	0.903	0.983	1.000	0.500	0.647	1.000	0.729	0.800	1.000	1.000	0.917
Rec	0.983	0.596	0.937	1.000	1.000	0.887	0.889	0.879	1.000	1.000	1.000	1.000
Acc	0.977	0.993	0.998	1.000	0.999	0.988	1.000	0.992	0.999	1.000	1.000	1.000
F	0.987	0.718	0.959	1.000	0.667	0.748	0.941	0.797	0.889	1.000	1.000	0.957

7 Discussion

This paper covers several aspects: a prosodic labeling scheme, an automatic annotation system and its evaluation. Questions could be raised about each of these aspects: about the nature of prosodic labeling scheme (7.1), its adequacy (7.2), its use by human transcribers (7.3), the evaluation procedure, more precisely the validity of the reference corpus (7.4), and the data selection (7.5).

7.1 The comprehensive nature of the prosody annotation

A *comprehensive annotation of prosody* includes both tonal (i.e. pitch-related) aspects and a representation of prosodic structure (including the position of stressed syllables and prosodic boundaries). Here, the scope was deliberately restricted to pitch features of speech. The main reason for doing so is that a tonal annotation of the type proposed in this study can be obtained in a *language-independent* way. The algorithm relies almost exclusively on acoustic information in the speech signal itself. The successive processing steps (such as the segmentation into syllabic nuclei, pitch stylization and the final interpretation of the sequence of pitch targets and pitch movements) all rely on the processing of information calculated on the speech signal. As a result, the tonal annotation system may be used to annotate a wide variety of languages. (Whether this includes tone languages is still an open question.) In order to identify prosodic structure, however, it would be necessary to detect syllable stress and prosodic boundaries, and it is unclear whether this can be achieved without language-specific information, such as the position of word stress or language-dependent rules about the contribution of various prosodic features. In addition, the proposed tonal annotation may also be applied to languages for which prosodic structure has not yet been described.

As said above, the obtained tonal annotation may be integrated in a later stage with information about prosodic structure. For instance, for autosegmental representations, pitch levels will only be specified for syllables in certain positions in prosodic structure, such as the syllables at a pitch accent, or at a prosodic boundary. So the procedure for tonal annotation would remain mostly unchanged, but the annotation would only be kept for particular syllables.

This approach is illustrated by figure 7. Tier 4 provides a prominence judgment (to be obtained manually or automatically); tier 5 shows the automatic tonal alignment generated by our system. A prosody model specifying pitch levels and major pitch movements, for prominent syllables as well as for the syllables immediately preceding or following them, results in the transcription of tier 6 (stress is indicated as in IPA). This only requires the detection of prominence and the suppression of tonal information in particular syllables. As we move to more abstract representations, the annotation becomes more concise. The autosegmental style annotation of the last tier is obtained by mapping the labels indicating pitch levels and movements to ToBI-like notation indicating pitch accent and tones (but not the pitch movements themselves). Obviously, this mapping will be language-specific and model-specific. Also, additional tones (e.g. boundary tones) should be added to this tier to adequately describe the pitch contour according to ToBI conventions.

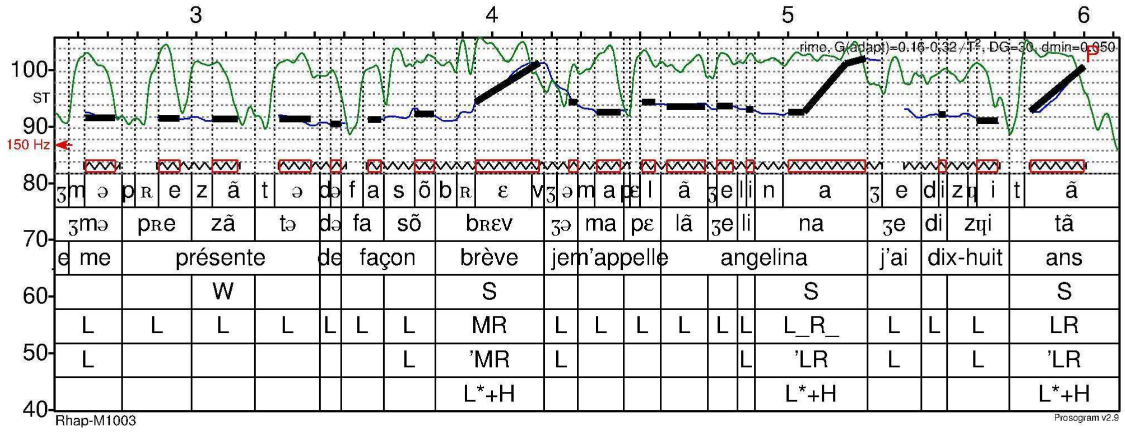

Fig. 7. Alternative tonal annotations for the excerpt of figure 5. The successive tiers show the phonetic alignment, the syllables, the text, a manual prominence annotation (W=weak, S=strong), the automatic

tonal labeling and two alternative labelings which may be derived automatically from the information in the prominence tier and the automatic tonal annotation tier. See text for further details.

7.2 The adequacy of the annotation convention

Whether this transcription provides an *adequate* representation of prosody obviously depends on its goal. The latter could be linguistic description of prosody, transcription of prosody in discourse, analysis of emotion in speech, text-to-speech synthesis, automatic annotation, and so on. Speech technology applications, for instance, require a representation which may be used in speech synthesis to generate the f0 specification of an utterance, and in analysis to automatically obtain the prosodic representation of an utterance. In linguistic description the goal is usually different. Still, the sheer number of competing representations of prosody demonstrates the lack of consensus on the choice of the annotation and hence on the issue of adequacy. To quote Hirst (2005, p. 335): “there is not even a real consensus on the way that prosody should be represented phonologically”. For these reasons the present system does not aim at an annotation according to some phonological model for a particular language, but rather at a transcription of pitch related features in speech, which may be computed in an automatic way, which may be applied to many languages, and which is based on psycho-acoustic observations of tonal perception.

It could be argued that an adequate transcription should be *linguistically meaningful*. But once again, there is little agreement on what this means. According to Taylor (2000, p. 1698) a prosodic representation is linguistically meaningful when it “contains information which is significant to the linguistic interpretation of an utterance’s intonation”. This view usually entails the following requirements: (1) the representation should be *constrained*, compact and without redundancy; (2) it should provide a *wide coverage* of prosodic phenomena and (3) *capture distinctions* which are perceptually different. Although the labeling used by Polytonia clearly meets the latter two requirements, it could be argued it is unsufficiently constrained. In our view, however, constraints should not be imposed *a priori* on the basis of theoretical or functional considerations (as is the case for most phonological models), but rather should result from prosodic features detected in the speech signal, such as prominence and pause, which will be integrated in a later stage of analysis.

7.3 The suitability of the annotation convention for human transcribers

The present study does not include an *experiment with human transcribers* to determine whether the labeling scheme may be used by human transcribers and with high inter-transcriber agreement. But the goal of this paper is not to propose a transcription of prosody suited for human transcribers, but rather to show (1) that a perceptually motivated tonal transcription can be computed automatically from the acoustic signal, without the need for a phonological model, (2) that this annotation is compatible with a more abstract (e.g. phonological) representation.

Actually, transcription experiments with human annotators are available only for a few labeling schemes, including ToBI (in its original form) and RaP (Dilley et al. 2006). For other major annotation conventions, such as INTSINT, no such experiments have been described to this date. The lack of such experiments does not imply the labeling scheme is necessarily invalid.

Also, the fact that subjects may be trained to use a particular labeling scheme does not necessarily imply it is adequate or linguistically meaningful, but only that subjects may be trained to use it in a consistent way.

Clearly, an evaluation of the consistency and inter-transcriber agreement of transcriptions by human coders using the Polytonia labeling convention would provide insight in the usefulness of the labeling scheme as a tool for manual annotation. However, given the size of such an experimental study, it would constitute the subject for an independent study.

It is well known that the judgments of human annotators are based not only on acoustic properties, but also on lexical and syntactic information, their knowledge of the language and their expectations. As a result, transcriptions may differ considerably between subjects. Automatic annotation, on the other hand, does not use these sources of information, but remains close to the acoustic facts.

7.4 The validity of the reference corpus

The evaluation procedure described in section 6 produces encouraging results. However, objections could be made with respect to the size and nature of the reference corpus, and the exclusion of syllables without detected f0.

A new type of tonal transcription is introduced in this paper, for which no reference corpus is available. As a result, a reference corpus had to be created and its size was determined mainly by time constraints, i.e. by the time-consuming nature of the manual validation of the reference corpus. This corpus has a total duration of 834s (13 min 54s). Whether this is *sufficiently large* is difficult to decide. Tables 5 and 6 show that results are very similar for subcorpora C1, C2 and C3, for female (e.g. C1) and male (e.g. C2) speakers, and for French (C1, C2) and Dutch (C3), except for symbols (pitch levels or movements) with a low frequency. (This low frequency is explained by the actual use and function of these prosodic forms in spontaneous speech.). These observations suggest that differences in speaker and language only marginally affect the results. Corpus size matters, of course, but other aspects are equally important, such as the number of speakers, their age difference, the inclusion of male and female speakers, the diversity of their voices, the inclusion of multiple languages (with quite different rhythmic structures), and the use of spontaneous speech rather than read speech. In this respect, the evaluation corpus scores rather high.

The reference corpus was obtained by having one expert correct by hand the output of the automatic annotation system. This approach has two important drawbacks: the small *number of judges*, and the *bias* introduced by using the system under evaluation for creating a reference.

If the obtained annotation were to match that of human transcribers, obviously the reference corpus would have to be based on the transcriptions of multiple judges. Instead, we are evaluating whether the annotation system detects tonal events which are perceptually distinct, such as pitch movements (rise, fall, level), simple and compound movements (e.g. rise vs. rise-fall), large vs. small movements, pitch levels (B, L, M, H, T), and so on. Earlier experiments using resynthesis of stylized pitch contours (Mertens et al., 1997) show pitch perception varies to some extent between listeners. As expected, subjects with musical training make finer pitch distinctions than the average listener in the subject group. By using a single highly-trained judge, in combination with repeated listening of small stretches of speech and inspection of the analysis output, clearly the resulting reference corpus will reflect the perception of someone with good pitch perception and it will outperform the perception of the average listener. As a result, the resulting reference corpus makes very high demands on the annotation system. And if the system is able to detect most tonal events in the reference corpus, it is likely that these will include those observed by the average subject.

The use of the output of the automatic annotation system to bootstrap the reference corpus introduces an important bias, since the judge will be inclined to accept the labels proposed by the system (unless they are clearly wrong). Unfortunaly, the creation of the reference corpus from scratch is so time consuming, that it is not feasible.

7.5 Syllables discarded from the evaluation data set

Some syllables in the reference corpus were *discarded from the evaluation data set*. This is the case *when f0 is not detected* (unvoiced sounds) or when other properties such as creak, hesitation or f0 discontinuities suggest that f0 measurement is likely to be unreliable. Our goal is to evaluate the tonal annotation system, not the f0 measurement on which it is based. f0 detection is a highly complicated matter (Hess, 1983), which lies outside the scope of this study. The preparation of the reference corpus clearly shows that the quality of the initial f0 measurement has a strong impact on the results obtained by the system, and that spontaneous speech and non-modal phonation pose many problems to standard f0 detection algorithms.

Syllables may also be eliminated in the evaluation of pitch level assignment, as mentioned earlier (6.2.1), *when pitch level is not assigned*. As pointed out above, this aspect is partly related to the identification of prosodic structure (position of stressed syllables and prosodic boundaries), which is beyond the scope of this study.

7.6 Interesting features of the annotation system

Despite the limitations discussed above, the *Polytonia* annotation system has several interesting properties.

First, it provides a very *narrow transcription* of tonal aspects, indicating pitch movements, their direction and size, as well as pitch level and pitch range (bottom, top). No other transcription provides this much detail when tonal aspects are concerned.

Second, the approach allows for a *speaker-independent* annotation of tonal features. The system handles speakers with normal, narrow or wide pitch range; it automatically adapts to the speaker's vocal range. Pitch range span and key are calculated on the speech corpus under analysis and various thresholds used by the system are adjusted to these values.

Third, the tonal annotation system is *language-independent*. It does not refer to properties of particular languages, such as part of speech information or syntactic features. As a result, the system may be applied to many languages, to obtain a tonal annotation for existing speech corpora. (Language-specific syllabic segmentation is automatically taken into account, as the segmentation into syllabic nuclei calculated by the system is based on the temporal alignment of sounds and syllables provided by the corpus annotation.)

Fourth, the system uses little information other than the *acoustic* signal itself. In this study the phonetic alignment was used to guide the segmentation into syllabic nuclei in order to obtain a precise evaluation of the tonal annotation, without the interference of segmentation errors. Many corpora used in prosody research already include an annotation of phonemes and syllables. Moreover, the system may also be applied using a fully automatic segmentation of the speech signal, resulting in an annotation tool which does not require any annotation whatsoever. In our view this property is crucial for the usefulness of the system. It is often claimed that prosody signals prominent information ("focus") as well as syntactic or discourse structure, which makes it useful to detect these prosodic events in the speech signal. However, if a detection system, as is often the case, requires lexical, morphological or syntactic information in order to identify these prosodic events, then it becomes somewhat circular, because the structural information which is obtained by the system is already available in the annotation it needs. The latter problem does not apply to our system.

Fifth, the approach described in this paper does not require a *training corpus*. This property constitutes a major advantage over common techniques for automatic classification by supervised learning, which all require training corpora. Since the validation of corpora (both training and reference corpora) is extremely time consuming (Tamburini and Caini, 2005, p. 34; Jeon and Liu, 2012, p. 446), the need of training corpora constitutes a major obstacle for the realization of automatic annotation systems for new prosodic transcriptions, for which such corpora are lacking. This obstacle does not apply to our system.

Finally, the transcription is "*theory-friendly*" (Hirst, 2005, 2011), because it is compatible with a number of theoretical approaches to the representation of tonal aspects in speech. It would be fairly straightforward to map the obtained tonal annotation to other annotation schemes, because the latter are generally less narrow. However, basic assumptions of the models may differ, complicating the conversion from one annotation into the other. For instance, one important theoretical choice made by our tonal transcription concerns the temporal alignment of pitch movements. The model assumes that the start of an intra-syllabic pitch movement is aligned with the vowel onset and that the movement is spread over the syllable rhyme. This assumption is rarely made elsewhere and needs to be examined in detail in experimental studies.

Our future research will focus on the detection of other aspects of prosody in continuous speech, including prominence, lengthening, stress and prosodic boundaries. The combination of these prosodic features with the tonal aspects will result in a more comprehensive transcription of prosody.

8 Conclusion

In this paper we have introduced a fine-grained notation for pitch-related aspects of speech and a described an automatic annotation system using this transcription.

The transcription assigns labels indicating pitch level and pitch movement to the sequence of syllables. It distinguishes 5 pitch levels: low (L), mid (M), high (H), top (T) and bottom (B). The first three are defined relatively and locally, on the basis of pitch changes in the immediate left context; the other two are defined relative to the observed global pitch range (vocal range) of the speaker. For pitch movements, the transcription indicates the direction (rise, fall, level) and the size of the pitch change, where the size categories are adjusted to the speaker's pitch range. Complex pitch movements are represented as sequences of simple ones.

This transcription is used by an automatic annotation system, which includes several processing steps: parameter calculation, segmentation into syllabic nuclei, pause detection, pitch stylization based on a model of tonal perception, pitch range estimation, analysis of intra-syllabic pitch variations, and pitch level assignment.

The automatic annotation system was evaluated on a reference corpus of 13 minutes of speech, including two female and three male speakers of French and Dutch. Values obtained for measures of precision, recall, F-measure and accuracy are very high for pitch levels and pitch movements for which sufficient data were available in the test corpus.

Reference List

Ananthakrishnan, S., & Narayanan, S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Trans. on Audio Speech and Language Proc.,* 16(1), 216-228.

Alessandro, C. d', & Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language,* 9(3), 257-288.

Bartkova, K., Delais-Roussarie, E., & Santiago-Vargas, F. (2012). ProsoTran: a tool to annotate prosodically non-standard data. Speech Prosody 2012.

Beckman, M.E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In Jun, S-A. (Ed.), *Prosodic Typology* (pp. 9-54). Oxford: Oxford University Press.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer [Computer program]. Version 5.3.10, retrieved 12 March 2012 from http://www.praat.org/

Braunschweiler, N. (2005). The Prosodizer - Automatic Prosodic Annotations of Speech Synthesis Databases. *Proceedings Speech Prosody* (Dresden).

Campione, E., Hirst, D., & Véronis, J. (2000). Automatic Stylisation and Modelling of French and Italian Intonation. In Botinis, A. (Ed.) *Intonation: Analysis, Modelling and Technology* (pp. 185-208). Dordrecht: Kluwer Academic Publishing.

Campione, E., & Véronis, J. (2001). Etiquetage prosodique semi-automatique des corpus oraux. *Actes TALN*, Tours, 2-5 juillet 2001.

Crystal, D. (1969). *Prosodic systems and intonation in English.* Cambridge: Cambridge University Press.

De Looze, C. & Hirst, D.J. (2010). Integrating changes of register into automatic intonation analysis. *Proceedings of the Speech Prosody 2010 Conference*. Chicago. 4 pages.

Dilley, L., Breen, M., Gibson, E., Bolivar, M., & Kraemer, J. (2006). A comparison of inter-coder reliability for two systems of prosodic transcriptions: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). *Proceedings of the International Conference on Spoken Language Processing*, Pittsburgh, PA.

Escudero, D., Aguilar, L., del Mar Vanrell, M., & Prieto, P. (2012). Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system. *Speech Communication,* 54, 566–582.

Geoffrois, E. (1995). *Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole*. Ph.D. Université Paris XI Orsay, 20 décembre 1995.

Grabe, E., Post, B., & Nolan, F. (2000). Modelling intonational Variation in English. The IViE System. In Puppel, S., & Demenko, G. (Eds.). *Proceedings of Prosody 2000.* Adam Mickiewitz University, Poznan, Poland. (2-5 October, 2000, Krakow, Poland.)

Grice, Martine (2006) Intonation. In Brown, Keith (ed.) *Encyclopedia of Language and Linguistics*, 2nd Edition. Elsevier: Oxford, vol 5, pp. 778-788.

Hart, J. 't (1998). Intonation in Dutch. In Hirst, D., & Di Cristo, A. (Eds), *Intonation systems: a survey of twenty languages* (pp. 96-111). Cambridge: Cambridge University Press.

Hart, J. 't, Collier, R., & Cohen, A. (1990). *A perceptual study of intonation.* Cambridge: Cambridge University Press. 227 pp.

Hermes, D. (2006). Stylization of pitch contours. In Sudhoff, S. et al. (Eds.), *Methods in Empirical Prosody Research* (pp. 29-61). Berlin: Walter de Gruyter.

Hess, W. (1983). *Pitch determination of speech signals. Algorithms and devices*. Berlin: Springer.

Hirst, D.J. (2005). Form and function in the representation of speech prosody. *Speech Communication,* 46, 334–347.

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., Nicolas, P.,& Espesser, R. (1991). Coding the F0 of a continuous text in French: An experimental approach. *Proc. International Congress of Phonetic Sciences,* Aix en Provence, France (1991), 234–237.

Hirst, D. J., & Di Cristo, A. (1998). A survey of intonation systems. In Hirst, D., & Di Cristo, A. (Eds.) *Intonation Systems. A Survey of Twenty Languages* (pp. 1-44.) Cambridge: Cambridge University Press.

Honorof, D. N., & Whalen, D. H. (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: Lund University Press.

House, D. (1995). The influence of silence on perceiving the preceding tonal contour. *Proc. Int. Congr. Phonetic Sciences* 13, vol. 1, 122-125. (Stockholm 1995)

House, D. (1996). Differential perception of tonal contours through the syllable. *Proceedings of International Conference of Spoken Language Processing*, 2048–2051. (Oct. 3-6, 1996. Philadelphia, PA, USA)

Jeon, J. H., & Liu, Y. (2012). Automatic prosodic event detection using a novel labeling and selection method in co-training. *Speech Communication, 54,* 445-458

Jun, S.-A. (Ed.) (2005). *Prosodic Typology.* Oxford: Oxford University Press.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustic Society of America, 118,* 1038-1054.

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

Ladd, D. R. (2008) *Intonational Phonology*. Cambridge: Cambridge University Press. Second edition.

Martin, Ph. (2009). *L'intonation du français.* Paris: Armand Colin. 256 pp.

Mertens, P. (1987a). *L'intonation du français. De la description linguistique à la reconnaissance automatique*. Unpublished Ph.D. (University of Leuven)

Mertens, P. (1987b). Automatic segmentation of speech into syllables. In Laver, J., & Jack, M.A. (Eds.) Proceedings of the European Conference on Speech Technology. Vol. II, 9-12. Edinburgh: CEP Consultants.

Mertens, P. (1989). Automatic recognition of intonation in French and Dutch. *Eurospeech, 89,* 1, 46-50.

Mertens, P. (2004a). The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In Bel, B.,& Marlien, I. (Eds.) *Proceedings of Speech Prosody 2004,* Nara (Japan), 23-26 March 2004.

Mertens, P. (2004b). Un outil pour la transcription de la prosodie dans les corpus oraux. *Traitement Automatique des langues, 45 (2),* 109-130.

Mertens, P., Beaugendre, F., & Alessandro, Ch. d' (1997). Comparing approaches to pitch contour stylization for speech synthesis. In Santen, J.P.H. van, Sproat, R. W., Olive, J. P., & Hirschberg, J. (Eds.) *Progress in Speech Synthesis* (pp 347-363). New York: Springer Verlag.

Rosenberg, A. (2010). AuToBI - A Tool for Automatic ToBI Annotation. *Proceedings Interspeech 2010*.

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., Di Cristo, A., Hirst, D., Martin, Ph., & Nishinuma, Y. (1981) *L'intonation. De l'acoustique à la sémantique.* Paris: Klincksieck. 364 pp.

Silverman, K., Beckman, M., Pitrelli, M., Ostendorf, M., Wightman, C. , & Price, P. (1992). TOBI: a standard for labeling English prosody. *Int. Conf. on Spoken Language Systems,* 867-870.

Smalley, W. A. (1964). *Manual of Articulatory Phonetics*. New York: Practical Anthropology, 512 pp.

Tamburini, F., & Caini, C. (2005). An automatic system for detecting prosodic prominence in American English. *International Journal of Speech Technology* 8(1), 33-44.

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, 107(3), 1697-1714.

Wagner, A. (2009). Analysis and recognition of accentual patterns. *Proceedings Interspeech* 2009 (6-10 Sept., Brighton, UK).

Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Trans Speech and Audio Processing, 2,* 469-481.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication,* 46, 220-251.