10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

ToBI Representations in Intonational Phonology: Time for a (melodic) change?

Philippe Martin

LLF, UFRL, Université de Paris, France
philippe.martin@linguist.univ-paris-diderot.fr

Abstract

The ToBI notation system has been the dominant transcription system used in the Autosegmental-Metrical framework. This system, among other drawbacks, doesn't really integrate listener tone perception, neither gives a clear account of the prosodic structure function in the sentence. It mixes frequently phonology and phonetics by being too close to the fundamental frequency curves obtained by acoustic analysis.

Therefore, an alternate phonological representation of sentence intonation should address the following points:

1. Instead of using ToBI High and Low tone targets, the perception of melodic variations could be better approximated using the rate of melodic change (glissando). This would allow to differentiate pitch changes which are perceived from those which are perceived as static tones.
2. Allowing accent phrases pitch accents to interact would allow to give a proper account for the incremental temporal aspect of the prosodic structure generated by the speaker, which does not result from some global mechanism involving the whole sentence at once. This leads to consider pitch accents as markers of dependency relations between accent phrases.

Describing accent phrases pitch accents as melodic contours reveals them as markers of AP's hierarchical sub-structures inside IP's ended by complex melodic contours merging pitch accent and boundary tone.

Index Terms: prosodic structure, ToBI, dependency grammar, glissando, melodic contour.

1. Introduction

Appeared in the 1980's, and becoming a standard during the 1990-2000 period, the ToBI notation system, basically transcribing high and low tone targets, has been the de facto standard in phonological representation of sentence intonation. Originally designed for American English [1], its extensions and adaptations to many other languages silenced other notation systems which became in practice completely invisible to researchers in prosody.

However, many problems can be identified pertaining, not to the principles underlying the ToBI notation, but towards its use. While the apparent difficulty to transcribe a moderate vs. a large fundamental frequency span can be remedied contrasting H and HH for example (or L and LL for low tone targets), its actual use linked to a direct F0 curve observation can be problematic. Leaving aside the sometimes-questionable reliability of pitch trackers, many transcriptions do not hesitate to considered insignificant pitch changes as low as 5 Hz as high tone targets, therefore transcribed as a H tone, the same notation than a 45 Hz rise for example [2], the duration of the pitch change being totally absent in the annotation.

Another problem is linked to the frequent confusion between phonetic and phonological transcription. Even if the goal of phonological transcription is to give a proper account of whatever prosodic events results from the speaker control, opposed to what is contingent to physiological constrain which pertains to phonetics, it should not be forgotten that the observed prosodic data result from the superposition of many processes. A sound representation system should clearly define from the beginning what specific function is addressed when a transcription system is used. More than often, current phonological description of sentence intonation seems to address many functions at once, resulting in total confusion in the validity of the description by mixing non-pertinent characteristics of prosodic events related to multiple roles.

2. Another viewpoint

Perhaps the first thing to do is to reconsider the relationship of sentence intonation with the other linguistic objects in the sentence. Roughly, the Autosegmental-Metrical model (AM) envisions the prosodic events, and particularly the prosodic structure, as an emanation of syntax. Accent phrases (AP), as minimal prosodic units, are assumed to contain only one non-emphatic stressed syllable and therefore one content word (verb, noun, adjective or adverb, opposed to lexical words, conjunction, pronoun, article, preposition…). Their (optional) assembly into intermediate intonational phrases (ip), then of ip into Intonational Phrases (IP), and finally of IP into the sentence prosodic structure (PS) proceeds a priori with congruence with syntax, the occasional non-congruence being described relatively to the sentence syntactic structure considered as the reference in the process.

However, the theoretical view giving a reference status to syntax should be reevaluated at the light of recent experimental findings. In this respect, the emergence of a large amount of data coming from spontaneous speech analysis could be fruitful. More and more arguments, instead of considering the prosodic structure as the cherry on the syntactic cake, suggest that the sequences of prosodic events indicating a (partial) PS appear to be planned before the AP's syntactic chunks [3]. The analysis of hesitations, abandons, reformulations and repetitions in spontaneous speech production shows clearly that the corrected AP's segments do not involve any prosodic change from the planned ones, whose melodic markers remain in place [4]. This apparently applies to observed error analysis occurring in read as well as spontaneous speech.

The other possible theoretical revision concerns the annotations system itself. Other than not being very intuitive [5], transcription by tonal targets, especially by a unique level tonal target high or low, is not really backed by experimental data pertaining to speech perception (at least for non-tonal

10.21437/SpeechProsody.2020-37

languages). Even pioneer work such as conducted by the IPO labs in the seventies shows that human perception proceeds by pitch change and not by pitch levels [6]. Rather than tonal targets, a notation operating on the base of pitch contours may reveal much more satisfactory from both points of view, perceptual and experimental. The use of a perception criterium such as the glissando threshold, however approximative, gives some indication about the pertinence of F0 movements observed on pitch tracks, separating pitch changes perceived from those which are perceived as static tones.

Abandoning the notion of tone targets, implicitly derived from the western musical notation as most musical instruments produce static tones and not pitch changes, one can be closer to the reality of prosodic events in speech, which are manifested by changes and not at all by static tones.

3. Phrasing

Let's define the prosodic structure as a hierarchical organization of accent phrases, the minimal prosodic units, leaving aside the Strict Layer Hypothesis [7]. This means that accent phrases can be assembled in multiple levels, all of the same kind, i.e. as sequences of AP's. In languages such as English or Italian, content words bear a morphological syllabic stress, essentially located on its stressed syllable vowel (following voiced consonants will be left apart). It follows from the definition that if an accent phrase contains a content word, no other word can be stressed (except an eventual emphatic stress). But it does not necessarily say that a grammatical word cannot be stressed and be part of an accent phrase where no other word would be stressed. In a sequence such that *I believe that, for the most part,…*the conjunction *that* can be stressed and constitute an accent phrase in itself: *I bel**ie**ve th**a**t, for the m**o**st p**a**rt,…*.

Non-lexically stressed languages such as French or Korean may be considered to see things more clearly. Although it may be surprising or even unbelievable to some linguists, French has a stress system which does not indicate any morphological word boundary, and places a stress on any word last syllable, subject to rhythmic constrains. These constrains specify that no consecutive stressed syllables cannot be closer than about 250 ms, and cannot be separated by more than some 1250 ms. Incidentally, these limits established experimentally appear to be governed by the range of frequency variations of the delta brain oscillations [8].

The stress rule in French consists to assign an AP stress to the final syllable of any word, provided the spacing limits between consecutive stressed syllables is respected, and provided the resulting accent phrase corresponds to an entry in the lexicon, not of words, but of accent phrases. This implies that an accent phrase in French can contain more than one content word, depending on the speaker speech rate, or no content word at all.

An eurhythmicity process plays also a role in balancing the duration of consecutive accent phrases, either by adjusting the speech rate from one accent phrase to the next (spontaneous speech), or by selecting the words contained in accent phrases to attain a comparable number of syllables (red speech) [8].

4. Melodic contours as primitives

The next step consists of defining appropriate natural classes of prosodic events to give a satisfactory account of the function selected, to indicate the prosodic structure itself or another role given to sentence intonation such as its modality.

The location of these events is already known: they should be located on stressed syllables, and possibly on another remarkable location such as the last syllable of accent phrases. In French, these positions are the same as the accent phrase last syllable is necessarily stressed.

Instead of adopting High and Low as tonal target primitives, in order to follow the melodic curve more closely while not being abused by insignificant changes in F0, the basic contours first feature will be Falling and Rising. To differentiate contours perceived as melodic changes from those perceived as static tones, the glissando threshold constitutes a satisfactory possibility, even if it constitutes an approximation, separating F0 perceived change from those perceived as static.

Glissando values are computed with the formula (st2-st1)/(t2-t1) with st1 and st2 being respectively the semitones values at the beginning and at the end of F0 variation, (assumed to be sufficiently linear) and t2-t1 the duration of the pitch movement (semitones are given by st = 12 * log(F0/100.0) / log(2)). Values above a glissando threshold are considered perceived as melodic variations, those below as static tones, at 2/3 of the pitch change. Correction for concordant intensity change can also be included [9].

At this point we have 4 classes of contours: Rising above the glissando threshold, falling above the glissando threshold, and rising and falling below GL. These last categories can merge into one, called neutralized, as their melodic change rising or falling is not perceived as such.

Still we need at least two other classes, to differentiate terminal conclusive contours declarative and interrogative, as the basic categories of sentence modality. The rising (interrogative) and falling (declarative) melodic changes, which can be found on the last stressed syllable of the sentence, but also in certain cases on its final syllable (e.g. in Italian [10]), indicate the basic modalities declarative or interrogative of the sentence. Variants implying non-linear F0 variation are correlated with déclarative implicative, imperative for the declarative case, and doubt and surprise for the interrogative case [8]. Leaving these variants aside here, we have to select some more acoustic features to differentiate the terminal conclusive contours from the others.

The range of fundamental frequency variation is plausible option, leading to set the interrogative terminal contour with a larger span than the rising contour above the glissando threshold but not in final position in the sentence. Likewise, the declarative terminal contour, usually above the glissando threshold, could be characterized by the lowest F0 value in the sentence, traditionally described as a pertinent feature in experimental phonetics [11].

In summary, the classes needed for the phonological transcription of French sentence intonation are:

Cdec: terminal conclusive declarative contour, falling, reaching the lowest F0 level, above the glissando threshold.

Cint: terminal conclusive declarative contour, falling, reaching the highest F0 level, above the glissando threshold.

Cris: non-terminal rising higher than the glissando level.

Cfal: non-terminal falling higher than the glissando level.

Cneu: non-terminal rising or falling, lower than the glissando level.

The single binary +/- Extreme feature can be selected to differentiate Cdec from Cfal and Cint from Cris. Cris and Cfal contours are then differentiated with the +/- Glissando feature, Cris +Glissando, Cfal +Glissando, Cneu -Glissando.

Non-tonal lexically stressed languages need another class of contours called complex contour Ccon. A complex contour has two parts: falling in the stressed syllable and rising on the last syllable. If the last syllable of an AP, usually at the end of an Intonation Phrase IP, the melodic movement is also complex, rising on about 1/3 of the stressed vowel, and rising on the remaining part (and possibly on the adjacent voiced consonant if any).

5. Combinatorial prosodic structure

In order to establish the distribution of the melodic contours defined above, we proceed by increasing complexity of the prosodic structure. One of the points rarely evoked in AM phonological intonative description pertains to the possible neutralization processes from the configuration of a given prosodic structure.

A simple example involving a sentence with a single accent phrase will illustrate this point. In such a configuration, there is only one stressed syllable (and thus one stressed vowel) carrying a terminal conclusive contour, which can be declarative or interrogative (leaving the variants aside). Considering the contrast necessary and sufficient to effectively indicate this modality in this simple configuration, only one feature would be necessary, such as rising (marked) vs. falling (unmarked): Cdec = -Rising vs. Cint = +Rising. In a sentence with a single accent phrase, no other contrast is necessary to indicate its modality.

As the prosodic structure becomes more complex (at this stage alignment with syntactic chunks is not involved), the need for more contrasting features emerges. A PS with 2 accent phrases could have the first or the second AP or both ended by a conclusive terminal declarative contour Cdec (Cint left aside at this stage):

[Cdec Cdec] (1)

[Cx Cdec] (2)

[Cdec Cy] (3)

The first case implies either two distinct sentences, ended with a terminal conclusive contour, or a configuration with to prosodic ends associated to a single sentence. This latter possibility is called in French *complement différé* [12], where two (or more) independent PS are associated with a single sentence text, implying the explicit syntactic link *which* as in *I know many buildings* Cdec *which must be demolished* Cdec.

For case (2), the so far unspecified contour Cx must contrast with Cdec (otherwise we would fall back on case 1). Choosing among the classes defined above, it could be Cris, Cfal or Cneu. Cfal could be eliminated as rarely attested in the experimental data (except in specific speaking styles such as dictation). This leaves Cris Cdec and Cneu Cdec as well-formed sequences. The contrast between Cris and Cdec is ensured by the feature Extreme: Cris -Extreme, Cdec +Extreme.

From examination of data, the third case (3) corresponds to the Rheme-Theme configuration, characterized by a Cneu terminal contour placed after the terminal conclusive Cdec Cneu. The contrast Cdec Cneu is secured by the glissando threshold condition: +/- Glissando.

In summary, in a 2 AP's prosodic structure, melodic contour features necessary and sufficient on both paradigmatic and syntagmatic axis are shown in table 1 (paradigmatic contrasts are in parenthesis).

Table 1. *Necessary and sufficient features in a 2 AP configuration*

Cdec	Cdec (1)
(+Extreme)	(+Extreme)
(-Rising)	(-Rising)
(+Glissando)	(+Glissando)
Cx	Cdec (2)
-Extreme	+Extreme
+Rising	-Rising
(+Glissando)	(+Glissando)
Cdec	Cy (3)
+Extreme	-Extreme
-Rising	+/-Rising
+Glissando	-Glissando

From this table, we conclude that, Cx = Cris for (2), and Cy = Cneu for (3).

Prosodic structure with 3 accent phrases will of course involve more features ensuring more contrasts between the AP's melodic contours. Considering only PS ended by a declarative terminal conclusive contour, we have 3 possible configurations:

[Cx Cx Cdec] (4)

[[Cy Cx] Cdec] (5)

[Cx [Cy Cdec]] (6)

Melodic contour features necessary and sufficient to differentiate the 3 configurations involving 3 accent phrases are shown in table 2.

Table 2. *Necessary and sufficient features in a 3 AP configuration*

[Cx	Cx	Cdec (4)
-Extreme	-Extreme	+Extreme
+/-Rising	+/-Rising	-Rising
+/-Glissando	+/-Glissando	+Glissando
[Cx	[Cy	Cdec]] (5)
-Extreme	-Extreme	+Extreme
+Rising	+/-Rising	-Rising
+Glissando	-Glissando	+Glissando
[[Cx	Cy]	Cdec] (6)
-Extreme	-Extreme	+Extreme
-Rising	+Rising	-Rising
+Glissando	+Glissando	+Glissando

In conclusion: for (4) Cx can be Cris or Cneu, for (5) Cx = Cris, Cy = Cneu (no other solution), and for (6) Cx = Cfal, Cy = Cris.

An alternate solution for (6) is Cx = Cneu with -Glissando:

[[Cx	Cy]	Cdec] (6)
-Extreme	-Extreme	+Extreme
+/-Rising	+Rising	-Rising
-Glissando	+Glissando	+Glissando

6. Prosodic grammar

From the definition of melodic contours and their necessary and sufficient features as dependent of the complexity of the prosodic structure, the following set of rules can be considered for French:

Cdec -> Cdec Cdec (differed complement)

Cdec -> Cdec Cneu (Rheme-Theme)

Cdec -> {Cneu, Cris} Cdec (2 AP's)

Cris -> {Cneu, Cfal} Cris (in 3 AP's configuration)

Cfal -> Cneu Cfal

It can be also shown that the prosodic structure is defined by dependency relations "to the right" existing between melodic contours, as follows [4, 8]:

A neutralized contour depends on the presence later in the sentence of either a neutral, falling, rising or terminal contour.

Cneu -> {Cneu, Cfal, Cris, Cdec}

A falling contour depends on the presence later in the sentence of either a falling, or a rising contour.

Cfal -> {Cfal, Cris}

A rising contour depends on the presence later in the sentence of either a rising, or terminal contour.

Cris -> {Cris, Cdec}

7. Two examples

Fig. 1 gives an example of the melodic contours and the resulting prosodic structure on a read sentence in French: *deux alpinistes allemands ont trouvé le cadavre d'un homme dans un glacier* « two German mountaineers found the corpse of a man in a glacier ».

.

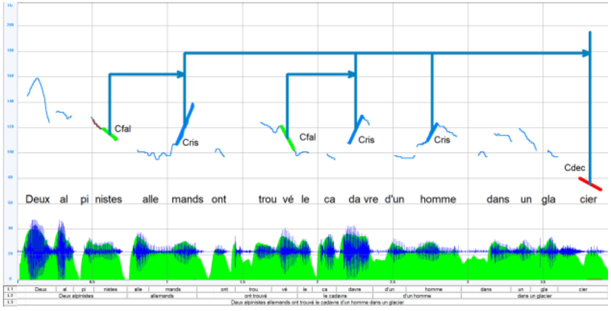

Figure 1. Annotated melodic contours and resulting prosodic structure of an example in French: *deux alpinistes allemands ont trouvé le cadavre d'un homme dans un glacier* « two German mountaineers found the corpse of a man in a glacier ».

The stressed syllables of the example determine the segmentation into the following accent phrases: [*deux alpin**is**tes*] [*allem**an**ds*] [*ont trouv**é***] [*le cad**a**vre*] [*d'un h**o**mme*] [*dans un glacier*]. The melodic contours derived from the fundamental frequency curve on stressed vowels are [*deux alpin**i**stes* Cfal] [*allem**an**ds* Cris] [*ont trouv**é*** Cfal] [*le cad**a**vre* Cris] [*d'un h**o**mme* Cris] [*dans un glacier* Cdec], which indicate the following prosodic groups:

[*deux alpin**i**stes* Cfal] [*allem**an**ds* Cris]
[*ont trouv**é*** Cfal] [*le cad**a**vre* Cris]
[*d'un h**o**mme* Cris]
[*dans un glacier* Cdec].

The resulting prosodic structure is therefore not congruent with the VP syntax (*ont trouvé*) (*le cadavre d'un homme*), which can be easily explained by the reading process operation from left to right…

8. Discussion

The glissando threshold, as the formula giving its value involves a parameter varying in the literature from 0.16 to 0.32 [16], but also as the estimation of change in F0 perception implies linearity of the melodic change. In reality, in most speech styles more or less close to reading style, the change in F0 is linear in both linear and logarithmic scale. However, concave or convex shapes can be found in some specific speech styles or sociogeographic productions (e.g. Swiss French).

9. Conclusions

Facing the dominance of the autosegmental-metrical model and the concomitant use of the ToBI transcription / notation system, we propose an alternate model radically different in terms of overall comprehension of the linguistic aspects of sentence intonation. Indeed, in the view presented here, the intonational grammar does not result from a collection of well-formed sequences of tone targets, where only boundary tones have a linguistic role by signaling the end of so-called Intonation Phrases as main assembly of accent phrases. On the contrary, they stem from a simple and coherent view of the sentence prosodic structure actually preexisting to syntactic units and organization.

The resulting grammar of intonation is then easily obtained from logical consideration pertaining to combinatorial analysis of prosodic structures, where the neutralization process plays a central role. Describing the relations between melodic contours assigned to accent phrases pitch accents leads to a dependency grammar based on the melodic contour's distribution, as observed on experimental data, which, contrary to most AM studies, go far beyond simple sentences with few words, and which is easily extended to spontaneous speech intonational analysis [4].

Furthermore, instead of tone targets aligned on IP boundaries and pitch accent having no interaction with each other, phonologically pertinent prosodic events are described in terms of perceived melodic changes, located on stressed syllables vowels, leading to classes of melodic contours experimentally easy to characterize: rising and falling, above or below the glissando threshold.

10. References

[1] Pierrehumbert, J. "The phonology and phonetics of English intonation", Ph.D. Thesis, Massachusetts Institute of Technology, Dept. of Linguistics and Philosophy, 1980.

[2] C. Gussenhoven, Transcription of Dutch Intonation, in "Prosodic Typology: The Phonology of Intonation and Phrasing", Sun-Ah Jun ed., Oxford University Press, 2005.

[3] P. Keating and S. Shattuck-Hufnage, "A Prosodic View of Word Form Encoding for Speech Production", *UCLA Working Papers in Phonetics*, 2002, 101: 112-156.

[4] Ph. Martin, *The Structure of Spoken Language. Intonation in Romance*, Cambridge: Cambridge University Press, 2015, 292 p.

[5] R. S. Schaefer et al. "Intuitive visualizations of pitch and loudness in speech" *Psychonomic bulletin & review*, vol. 23, 2, 2016, pp. 548-55.

[6] J. 't Hart, R. Collier & A. Cohen, *A perceptual study of intonation: an experimental-phonetic approach to speech melody*, Cambridge: Cambridge University Press, 1990, pp. xv + 212.

[7] E. O. Selkirk, "On prosodic structure and its relation to syntactic structure", in T. Fretheim, ed., *Nordic Prosody II*. Trondheim: TAPIR, 1978, pp.111-140.

[8] Ph. Martin, *Intonation, structure prosodique et ondes cérébrales*, London : ISTE, 2018.

[9] M. Rossi, "Le seuil de glissando ou seuil de perception des variations tonales pour la parole", *Phonetica* (23), 1971, pp. 1-33.

[10] M. Savino, "The intonation of polar questions in Italian: Where is the rise?", *Journal of the International Phonetic Association*, Vol. 42, 1, 2012, pp. 23-48.

[11] W. E. Cooper and J. M. Sorensen, *Fundamental frequency in sentence production*, New-York: Springer Verlag, 2012.

[12] Ch. Bally, "Intonation et syntaxe", *Cahiers de Ferdinand de Saussure 1*, 1941, pp. 33-42.