Phonetic representations for intonation

Gérard Bailly
Véronique Aubergé

1 Introduction

Besides its intrinsic ability to help the listener in segmenting utterances into linguistically-relevant parts of speech, intonation is an essential means for signaling *"...of how we feel about what we say, or how we feel when we say."* [Bol89, p.1]. How the cognitive representation of the utterance is encoded in the speech signal is still an open question. Two main approaches may contribute to our understanding of how intonation contributes to both the capture of the overall meaning of the message and the speaker's position *vis-à-vis* its own discourse: (a) a bottom-up approach aims to extract salient prosodic events with no linguistic a-priori. Such tentative approaches, linking these events to phonological constructs, face the problem of automatic extraction [Mer93, HNE91], the perceptual relevance of these events [tCC90, HR94] and the coherence of labelling [SBP$^+$92, HdCng], (b) a top-down approach aims to extract prosodic parameters that have sufficient statistical significance to carry given - contrastive - linguistic or paralinguistic tasks. This approach requires a careful design of large corpora. The resulting utterances should enable statistical analysis to extract significant prosodic correlates of underlying contrasts built in the corpus design. Careful design of the corpus and use of adequate statistical tools are the key problems of this top-down approach.

2 Building phonological representations with bottom-up analysis

For the last twenty years, specific models have suggested phonetic constructs which could be the elementary contrastive elements on which phonological representations can be build. These phonetic models essentially concentrate on fundamental frequency (F0). They incorporate global and local elements. Global elements are handled mostly at the utterance-level.

They include such constructs as declination lines [Pie79, Pie81, tHC73] or tonal grids [Gr91]. These rigid templates could be replaced by a more flexible "down-stepping" and pitch reset (see contribution by Kohler in this section). Local elements code an F0 curve into a sequence of pitch targets [Gr91, Hir83] connected by various interpolation functions, or pitch-pulses [FS71]. From a phonological perspective, these local pitch movements together with rhythmic patterns (pauses, syllabic lengthening) delimit phonological units such as accent, foot, prosodic phrase or intonation phrase. Generative phonology thus offers a way to describe how these phonological units interact to produce acceptable language-specific intonation [PB88] and how it relates to linguistic descriptions [Sel84]. These constituents are assumed to also define the rhythmic structure of the utterance[1]. It has been shown that the rhythmic structure derived by psycho-linguistic experiments [GG83, MG93] may carry fine details of the linguistic complexity [WSHOP92], as emphasised by Hirst [Hir93, p.17]: *"Recent corpus-based studies of durational cues for prosodic constituency have suggested that more than two levels of intonation units can be systematically distinguished."*. A solution consists of augmenting the abstract phonological representations with numerical values such as break indices [SBP+92] or stress levels (see kohler, this volume). In order to relate the variability of concrete phonetic representations describing pitch levels or movements - due to speech rate, speaking styles or dialects - to these abstract phonological representations, different proposals have been made: Hirst [Hir91] proposes to link the elementary tonal segments to high order phonological constituents whereas Gussenhoven [Gus91] argues that the intonational domain of a pitch accent cannot be linked to any of the constituents and introduces the *intonation domain* for each accent as "*...an independent constituent over which a single intonation contour spreads.*" (see marsi et al, this volume). These adaptations of surface phonological representations of the pitch curve and generation models are a indication that tonal segments are not independent and are often part of larger prosodic *contours* assuming different cognitive functions. Some stereotyped *melodic tunes* composed of a series of sustained pitches [Lad78] called also *cliché* tunes [FBF84] have been described in the litterature. These tunes seem to function as a whole favoring an holistic analysis [HY91].

[1] Note that so-called rhythmic structure mentioned above is often limited to the structural dominance within the utterance between accented syllables. Recent research [Cam92, BB94, Pas92] has shown that syllabic durations are organised in larger contrastive units within which gradual lengthening may occur.

3 Building phonetic models with top-down analysis

Another approach, introduced more recently in the field of speech synthesis, is to build and record large corpora based on strong linguistic or paralinguistic constraints. Statistical analysis can then be applied in order to capture prosodic regularities which arise from the implicit instructions - e.g. reading texts designed to cover the statistical range [Aub92, Tra92] - or by explicit instructions given to the speaker - e.g. different speaking styles for the same unmarked text. In this section Abe presents such a contrastive analysis. The same text is pronounced with three different speaking styles. Statistical analysis shows that styles differ in many dimensions including melodic contours, phonation/silence ratios ... but also vocalic quality such as spectral tilt or formant trajectories.

Thus direct links between prosodic parameters and cognitive representation of the utterance controlled or biased by instructions can be established. More global phonetic models are implied by this approach: the kinematics of the F0 curve is thus globally - and often directly - controlled by cognitive representations. Generation models with non-linear dynamics [Sag90, Tra92, LF86] or superposition models [Aub92, ;Abe, this volume] then convert cognitive representations into prosodic movements.

4 Phonetic representations and cognitive functions

Both approaches may contribute to proposals and assessments of intermediate phonological representations which elucidate the real contrastive phonetic representations. We still lack significant psychoacoustic and perceptual data on these degrees-of-freedom such as basic psycho-acoustic difference limens or perception thresholds [HR94], synchronisation of prosodic trajectories [Koh90] (see for example the $\pm early$ and $\pm late$ distinctive features of the intonation component presented in the paper by Kohler .

The question of how high-level phonological structures are encoded into prosodic trajectories is crucial for understanding how intonation is produced and identified in the communication process. In her introduction to the symposium "Prosody in situations of communication" at the XIIe International Congress of Phonetic Sciences, Cutler notes that *"Two of the major uses of prosodic information in situations of communication are to encode salience and segmentation."* [CN91, p.264]. In this section Marsi and colleagues study the influence of syntactic and phonotactic constraints on prosody as a cue for textual cohesion. Horne and Filipsson present a computational framework for deriving prosodic structure using both syntactic and new/given information: complementary to other linguistic levels such as syntax, prosody contributes to "highlight" new information.

This suggests that listeners are able to clearly distinguish between prominence in accordance with "normal" linguistic structure and prominence which emphasises speech segments. Consequently listeners appear to have some expectations on what is a "normally" structured intonation for a given message and capture as salient something emerging from this basis or something that deviates from an expected pattern.

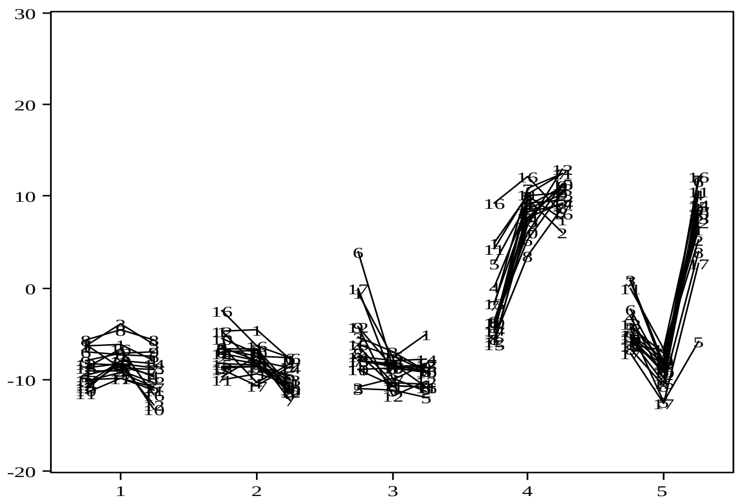

FIGURE 1. Superposition of melodic contours for utterances with five syllables typical of a incredulous interrogation [FBF84, p.181]. The "accent" on the penultimate syllable is part of a global contour resisting to variation of the syntactic and phonotactic substrate (from [MBA95]).

5 Prosodic prototypes

Global approaches have been proposed previously in the literature. Fónagy states for example that:

"The opinion that any utterance is an instantaneous production . . . applies only to the domain of modern poetry. Except for this limited domain, our utterances are made of large prestored pieces. . . " (Translated from [FBF84, p.182])

Books written by Fónagy [Fn82] and Bolinger [Bol81] are illustrated by multiple examples of such global intonational and rhythmic patterns as presented in Figure 1. Does this suggest that such global patterns are instances of prosodic *prototypes* which are the real basic elements of a prosodic *lexicon*? More experimental and psycho-linguistic work is needed to assess such proposals. Some authors incorporate global contours in their phonetic model of melody. Eva Gårding models the F0 as a superposition of global "slow" movements representing sentence or phrase intonation [Gr91] with local movements pertaining to an accent or tone. The first level is controlled by tonal grids connected by *pivots*. The second level is controlled by *turning points*. The tonal grid is thus controlled in a more complex way than two declination lines. This distinction between global and local models is also presented in [AS92], where a simple superposition model sums two main contributions to the F0 curve: (a) a global model assigns average F0 value for minor phrases and (b) a local model assigns then F0 average values and slopes for each syllable within each phrase. Véronique Aubergé [Aub92] introduces a synthesis scheme where the various linguistic levels (sentence, proposition, syntactic groups and sub-groups) are encoded into global prosodic multi-parametric contours. Low-level contours modulate the higher-level contours. This latter approach differs from the preceding ones in the fact that no assumption is made a priori between levels in terms of rate of change (as for example the difference between cut-off frequencies of breath and word group commands in Fujisaki's model [FK88]: each syllable contributes to the encoding of each level and higher-levels can use abrupt changes in intonation or rhythmic structure to encode cognitive representations.

6 Conclusions

Following Fónagy [Fn82] and Bolinger [Bol81] more attention should be paid in the future to prosodic means used by speakers to indicate how they "adhere" to their own discourse in order to emphasise global functions of prosody in the communication process and their interactions with salience and segmentation. Papers presented in this section deal with different apsects of these interactions.

The links between generation and analysis models of intonation should also be reinforced in order to control efficient dialogs. Our assumption is that *holistic* analysis of prosodic representations could reinforce this link and explain why prosody can be memorised and identified precociously by the child, independently from any pre-linguistic structuration [KM82, LT77].

7 References

[AS92] Masanobu Abe and Hirokazu Sato. Two-stage f0 control model using syllable based f0 units. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 53–56, 1992.

[Aub92] Vronique Auberg. Developing a structured lexicon for synthesis of prosody. In G. Bailly and C. Benot, editors, *Talking Machines: Theories, Models and Designs*, pages 307–321. Elsevier B.V., 1992.

[BB94] Plnio Barbosa and Grard Bailly. Characterisation of rythmic patterns for text-to-speech synthesis. *Speech Communication*, 15:127–137, 1994.

[Bol81] Dwight Bolinger. Some intonation stereotypes in english. *Studia Phonetica*, 18:97–101, 1981.

[Bol89] Dwight Bolinger. *Intonation and its uses.* Edward Arnold, London, 1989.

[Cam92] W.Nick Campbell. *Multi-level Timing in Speech.* PhD thesis, University of Sussex, 1992.

[CN91] Anne Cutler and D. Norris. Prosody in situations of communication: salience and segmentation. In *International Congress of Phonetic Sciences*, volume 1, pages 264–270, Aix-en-Provence, France, 1991.

[FBF84] Ivan Fnagy, Eva Brard, and Judith Fnagy. Clichs mlodiques. *Folia Linguistica*, 17:153–185, 1984.

[FK88] Hirose Fujisaki and H. Kawai. Realization of linguistic information in the voice fundamental f requency contour of the spoken japanese. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 663–666, 1988.

[Fn82] Ivan Fnagy. *Situation et signification. Prolgomnes un dictionnaire des noncs en situation.* Benjamins, Amsterdam, 1982.

[FS71] Hirose Fujisaki and H. Sudo. A generative model for the prosody of connected speech in japanase. *Annual Report of Engineering Research Institute*, 30:75–80, 1971.

[GG83] Jean-Pierre Gee and Franois Grosjean. Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:418–458, 1983.

[Gr91] Eva Grding. Intonation parameters in production and perception. In *International Congress of Phonetic Sciences*, volume 1, pages 300–304, Aix-en-Provence, France, 1991.

[Gus91] Carlos Gussenhoven. Tone segments in the intonation of dutch. In T.F. Shannon and J.P. Snapper, editors, *The Berkley conference on Dutch linguistics*, pages 139–155. University Press of America, Lanham, MD, 1991.

[HdCng] Daniel J. Hirst and Albert di Cristo. *Intonation systems: a survey of twenty languages.* Cambridge University Press, Cambridge, forthcoming.

[Hir83] Daniel Hirst. Structures and categories in prosodic representations. In A. Cutler and A. Ladd, editors, *Prosody: models and measurements*, pages 93–109. Springer-Verlag, Berlin, 1983.

[Hir91] Daniel Hirst. Intonation models: towards a third generation. In *International Congress of Phonetic Sciences*, volume 1, pages 305–310, Aix-en-Provence, France, 1991.

[Hir93] Daniel Hirst. Peak, boundary and cohesion characteristics of prosodic grouping. *Working Papers of Lund University*, 41:32–37, 1993.

[HNE91] Daniel Hirst, Pascale Nicolas, and Robert Espesser. Coding the f0 of a continuous text in french: an experimental approach. In *International Congress of Phonetic Sciences*, volume 5, pages 234–237, Aix-en-Provence, France, 1991.

[HR94] Dik J. Hermes and H. H. Rump. Perception of prominence in speech intonation induced by rising and falling pitch movements. *Journal of the Acoustical Society of America*, 96(1):83–92, 1994.

[HY91] Jill-Elaine House and Nicolas Youd. Stylised prosody in telephone information services : implications for synthesis. In *International Congress of Phonetic Sciences*, volume 5, pages 198–201, Aix-en-Provence, France, August 1991.

[KM82] P. K. Kuhl and J. D. Miller. Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception and Psychophysics*, 31:279–292, 1982.

[Koh90] Klaus J. Kohler. Macro and micro f0 in the synthesis of intonation. In J. Kingston and M.E. Beckman, editors, *Papers in Laboratory Phonology*, volume I, pages 115–138. Cambridge University Press, Cambridge, UK, 1990.

[Lad78] D. R. Ladd. Stylised intonation. *Language*, 54:517–541, 1978.

[LF86] A. Ljolje and Franck Fallside. Synthesis of natural sounding pitch contours in isolated utterances using hidden markov models. *TrASSP*, 34:1074–1080, 1986.

[LT77] C. N. Li and S. A. Thompson. The acquisition of tone in mandarin-speaking children. *Journal of Child Language*, 4:185–200, 1977.

[MBA95] Yann Morlec, Grard Bailly, and Vronique Auberg. Synthesis and evaluation of intonation with a superposition model. In *EUROSPEECH*, volume 3, pages 2043–2046, Madrid - Spain, 1995.

[Mer93] Piet Mertens. Intonational grouping, boundaries, and syntactic structure in french. *Working Papers of Lund University*, 41:156–159, 1993.

[MG93] Pascal Monnin and Franois Grosjean. Les structures de performance en franais : caractrisation et prdiction. *L'Anne Psychologique*, 93:9–30, 1993.

[Pas92] Valrie Pasdeloup. Dure inter-syllabique dans le groupe accentuel en franais. In *XIXe Journes d'Etudes sur la Parole*, pages 531–536, 1992.

[PB88] Janet Pierrehumbert and Mary Beckman. *Japanese tone structure*. MIT Press, Cambridge, MA, 1988.

[Pie79] Janet Pierrehumbert. The perception of fundamental frequency declinaison. *Journal of the Acoustical Society of America*, 66:363–369, 1979.

[Pie81] Janet Pierrehumbert. Synthetizing intonation. *Journal of the Acoustical Society of America*, 70(4):985–995, 1981.

[Sag90] Yoshinori Sagisaka. On the prediction of global fo shapes for japanese text-to-speech. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1:325–328, 1990.

[SBP+92] K. Silverman, Mary Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, Patty Price, Janet Pierrehumbert, and Julia Hirschberg. Tobi: a standard for labeling english prosody. *Proceedings of the International Conference on Spoken Language Processing*, 2:867–870, 1992.

[Sel84] E. Selkirk. *Phonology and syntax*. MIT Press, Cambridge, MA, 1984.

[tCC90] J. 'tHart, Ren Collier, and A. Cohen. *A perceptual study of intonation*. Cambridge University Press, Cambridge, UK, 1990.

[tHC73] J. t' Hart and Ren Collier. Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1:309–327, 1973.

[Tra92] Chris Traber. Fo generation with a database of natural fo patterns and with a neural network. In G. Bailly and C. Benot, editors, *Talking Machines: Theories, Models and Designs*, pages 287–304. Elsevier B.V., 1992.

[WSHOP92] C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P.J. Price. Segmental durations in the vicinity of prosodic boundaries. *Journal of the Acoustical Society of America*, 91(3):1707–1717, 1992.