INTONATION MODELS: TOWARDS A THIRD GENERATION.

Daniel Hirst

Institut de Phonétique d'Aix,
URA CNRS 261 Parole et langage
Aix en Provence, France

ABSTRACT

Intonation models up to now can be classified into two generations : single language descriptions and multi-language models. The next step will be the development of an integrated theory of intonation defining a number of independent levels of representation together with a specification of the relationship between these different levels. A sketch of a framework for such a theory is given as well as a number of questions we need to ask about the nature of the different levels of representation.

1. INTRODUCTION

All linguistic description is faced with the challenge of sorting out those facts which can be put down to some universal language faculty from those which are assumed to be specific to a given language. How do we decide, in other words, which facts are to be incorporated into a general model of language and which are to be analysed as language specific parameters of the model ?

The problem is particularly acute in the study of intonation which appears to be one of the most universal characteristics of human language and, paradoxically, at the same time one of the most specific characteristics of a given language or dialect.[17] The universal character of intonation is well established. A striking illustration of its language specific nature can be seen from the recent finding [24 and references there] that as early as 4 days after birth, babies appear to be capable of distinguishing recordings of their native language from recordings of other languages. The fact that similar results are obtained with low-pass filtered recordings but not with recordings played backwards suggests that such discrimination is based on prosodic information which can only be acquired during the pre-natal period.

I suggest in my contribution to this symposium that we can distinguish two generations so far in the history of intonation models, and that our knowledge of intonation has now reached the point where we can begin to envisage a model belonging to a third generation. In section 3, I outline what appear to me some of the desirable characteristics of such a model.

2. INTONATION MODELS.

2.1 Generation one: single language descriptions.

The first generation of intonation models consists of phonetic and/or acoustic descriptions of the intonation of particular languages. Probably the vast majority of research which has been carried out on intonation falls under this heading. Since such descriptions concentrate on the intonation of a single language, no principled distinction can be made between model and parameters. These descriptions, however, constitute the indispensable groundwork on which more general models can be built.

2.2 Generation two: multi-language models.

In the last fifteen years or so, a number of specific models have been proposed for the description of the intonation of several languages. Examples of models of this type are those which have been developped in Lund [10, 11] and Eindhoven [12, 13] as well as work in the generative phonology paradigm [28, 27, 14, 15, 21, 29].
These models differ from those of what I have called the first generation in that they explicitly tackle the problem of separating out universal characteristics of intonation systems, which are directly incorporated into the model itself, from language specific characteristics which constitute the parameters of the model. In addition, second generation models have a number of common characteristics. Firstly the models all incorporate a number of constructs (**tonal grid, declination line, focus marker, boundary tone** etc.) specifically designed for the description of intonation: Secondly, the various models are generally oriented towards a specific descriptive level - phonological, physiological, perceptual etc. Finally, the descriptions proposed within the framework of the various models are in general underdetermined by the data. Given a model and a set of data, it often seems possible to account for the data in more than one way by choosing a different set of parameter values.

2.3 Generation three: ?

What would a third generation model look like ? The major distinction between such a model and those of the second generation would be the existence of general principles constraining the model and its applications. The primitive constructs of the model should thus be determined as far as possible by more general linguistic principles and the choice of parameters for a description should be fixed on the basis of a limited number of empirical questions which can be asked about the intonation system of a language. A complete third-generation model of intonation would, moreover, not be restricted to a particular descriptive level but should provide a number of different levels of representation, including at least phonological, phonetic and acoustic levels.

An explicit characterisation of these different levels of representation will be crucial to the development of a coherent body of theory. The following remarks raise a certain number of questions concerning these representations together with some very tentative answers.

3. LEVELS OF INTONATIONAL REPRESENTATIONS.

Few linguists today would question the fact that intonation forms part of a speaker's overall cognitive representation of an utterance. Such a high-level abstract representation can be assumed to contribute both to the overall meaning of an utterance and to the way in which the utterance is pronounced. At the other extreme, the acoustic parameters of fundamental frequency and intensity, together with the slightly more abstract parameter of segmental duration, constitute the physical dimensions in which intonation can be expressed. Between these two extremes - phonology and acoustics - lies the whole field of phonetics. Ohala has recently argued [25] that "There is no interface between between phonology and phonetics". Di Cristo and I have suggested [17] that this is because phonetics is itself an interface between the cognitive (what can be thought) and the physical (what can be said).

3.1 Phonetic representations

Given the hybrid nature of phonetics as the link between the cognitive and the physical, it would seem to follow logically that a phonetic representation can only be defined on the basis of a prior theory of phonological representations. In fact, however, there is a constant interaction between phonology and phonetics. The more we learn about phonetic processes, the more we can incorporate into our phonetic model and the more we may wish to

question traditional assumptions about phonological representations.

A number of techniques have been used for the generation of fundamental frequency for speech synthesis. Many of these techniques such as contour concatenation or neural networks leave unanswered the question of the nature of phonological representations. There are, however, at present at least two plausible candidates for a phonetic model of fundamental frequency curves.

The first of these, sometimes called the **Target and Transition Model**, assumes that an Fo curve is represented as a sequence of target points and that a (generally monotonic) interpolation function accounts for the portions of the curve between these target points [9, 10, 14, 15].

The second type which I shall call the **Pitch-Pulse Model** represents an Fo contour not as a sequence of pitch targets but rather as a sequence of instructions to raise or lower the pitch, the local pitch-movement being generally superposed on a more global movement. [26, 6]

It is obvious that at some level of abstraction, the two models are formally equivalent - any curve that can be generated as a sequence of targets can also be generated as a sequence of movements and vice versa.

On the level of speech production it has been argued [6] that a pitch-pulse model more adequately represents the actual physical mechanism underlying the generation of pitch contours in natural speech. Even this, however, does not guarantee that pitch contours are **mentally** represented as pitch changes. It is possible for example that a motor control mechanism such as a Forward Model [23, 1] is developped during the babbling stage, providing the speaker with a direct mental mapping between pitch targets and the impulses needed to generate the targets.

Evidence from acoustic modelling suggests that a model incorporating pitch targets more adequately accounts for the observed data than one generating pitch changes: in an experiment using controlled sentences [22] , relative peak levels were observed to be more highly correlated than were the corresponding rises or falls.

It has been claimed [4, 10, 17] that intonation systems generally make use of two distinct types of pitch levels : **relative** levels, determined with reference to the preceding pitch level, and **absolute** levels determined with reference to a wider context, perhaps even to an absolute speaker dependent value. While relative levels are easily coded in terms of pitch movements, absolute levels are less easily expressed in this way [27]

On the perceptual level, it has been claimed recently [20] that it is only during areas of spectral stability, as in a sustained vowel, that pitch patterns are interpreted as movement configurations rather than as pitch levels, but that elsewhere interpretation in terms of pitch levels is predominant.

It seems that much of the evidence points in favour of a mental representation of a pitch contour as a sequence of pitch targets even if this is not necessarily the form which serves as input to the pitch-producing mechanism in natural speech.

3.2 Surface phonological representations.

The surface phonological representation of a pitch curve can be assumed to consist of a sequence of phonetically interpretable symbols which can in turn be derived from a more abstract phonological representation. This would in many ways be the equivalent for intonation of the IPA transcription system for segmental phonology. An example of a first approximation of such a system is **INTSINT** [17], an INternational Transcription System for INTonation, which makes use of two types of symbols corresponding to the distinction between Absolute and Relative pitch levels mentioned above. Absolute pitch levels include :

↑ ↓
Top Bottom

as well as **Mid** which is assumed only to occur at the beginning of an Intonation Unit and is consequently unmarked. Relative pitch levels, defined with respect to the preceding pitch target include :

↗ ↘ ↱ ↳
Higher Lower Upstep Downstep

An application of this system to the F0 pattern of a continuous text in French [19] suggests the interesting result that all these targets can be defined with respect to three absolute speaker independent pitch levels corresponding to the speaker's mean Fo (Mid) and two levels (Top; Bottom) fixed at a half-octave interval respectively above and below the mean. It remains to be seen, however, whether similar results will be obtained for other speakers and other languages. It is also not clear how such a model can incorporate other factors such as variable pitch range [22].

3.3 Underlying phonological representations.

I suggested above that a surface phonological representation of intonation should have two characteristics. First, it should be phonetically interpretable, and secondly we should be able to derive it from a more abstract underlying phonological representation, the nature of which is obviously far from being clear. I have argued [15] that we should be guided by a general principle to the effect that all phonological primitives which are needed for the description of connected speech are needed for at least some languages to describe lexical contrasts. Thus for example stop aspiration is lexically distinctive in Hindi but not in English or French. English, however, unlike French, possesses a phonological rule adding aspiration to voiceless stops at the beginning of stressed syllables. The result is that in connected speech, aspiration can become distinctive in English (but not in French) as in the minimal pair : [ðɪspʰɔt] 'this port' ≠ [ðɪspɔt] 'this sport'.

If we compare the case of aspiration with that of word-stress we find a similar range of effects since word-stress is lexically distinctive in some languages (like Russian and English) but not in others (like French and Hungarian). The latter, however, contain rules assigning stress to the final (respectively initial) syllables of words so that stress can become distinctive in connected speech (cf minimal pairs in French : ['drapo'stbl] 'Drap possible' (possible sheet) ≠ [dra'po'stbl] 'Drapeau-cible' (target-flag). A similar argument can be made concerning tonal categories which, while they only need to be lexically specified for tone languages such as Chinese or Yoruba, remain available for the phonology of other languages in order to derive surface representations.

I mentioned above that as we learn more about phonetic processes we may well be led to modify our conception of underlying phonological representations. In the case of intonation, an explicit phonetic model suggests that tonal representations may be considerably sparser than in standard autosegmental accounts [7, 8]. I have suggested [14, 15, 16] that tonal segments are in fact linked not directly to vowels or syllables but rather to higher order phonological constituents as in Figure 1.

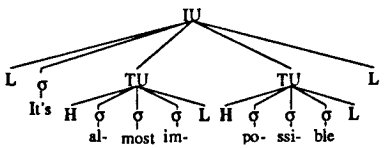


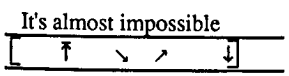
Figure 1 : Phonological representation for the sentence 'It's almost impossible'. **IU** = Intonation Unit, **TU** = Tonal Unit, σ = syllable, **H** = high tone, **L** = low tone.

A phonological structure of this type, removes the need for ad-hoc diacritic symbols indicating 'boundary tones' as proposed by Pierrehumbert [28]. Similar representations have since been proposed

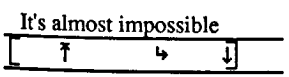
for a number of different languages including French [5], Swedish [2] and Japanese [27] as well as for an African tone language Kinyarwanda [3]. It seems probable that explicit acoustic modelling of other tone langauges could lead to similar results.

A representation of this type makes it possible to describe the variability found across the intonation patterns of different languages in terms of a small set of formal parameters. Thus if we compare the prosodic structure of English with that of French, we find that in English (as in Germanic languages in general), Tonal Units each contain one stressed syllable *followed* by an unlimited number of unstressed syllables, whereas in French (as probably in most Romance languages) Tonal Units contain one stressed syllable *preceded* by an unlimited number of stressed syllables. More formally Tonal Units can be described as *Left-Headed* in English and *Right-Headed* in French. The two languages also seem to differ in the sequence of tonal segments associated with Tonal Units in underlying representations. In English, each Tonal Unit contains the sequence [H L] as in Figure 1, whereas in French the underlying sequence seems to be [L H] [5]. Evidence from other languages [17] seems to suggest that the two parameters are independent.

The underlying representation in Figure 1 corresponds directly to the surface intonation pattern described as typical for many American English dialects which can be transcribed:



For most British English dialects, such a pattern is not at all typical. Instead a downstepping pattern is generally observed as in:



Both representations can be derived from the same underlying form illustrated in Figure 1 if we assume a further parameter which for British English allows just one single tonal segment to remain associated with each Tonal Unit except the last [15, 16].

The three parameters I have mentioned make it possible to generate an interesting variety of intonation patterns from otherwise similar underlying forms.

While it is obvious that the model I have sketched is far from constituting the Third Generation Model I evoked at the beginning of this discussion, I would argue that it possesses at least some of the characteristics.

4. CONCLUSIONS

I have presented a number of questions, together with some very tentative proposals concerning an integrated theory of intonation comprising several distinct levels of representation. My discussion of phonetic and acoustic representations has been exclusively concerned with fundamental frequency patterns although it is obvious that a complete description will need to account for variations of intensity and duration as well.

The aim of this presentation has been to sketch an overall framework such that at each step we can formulate our choice of representation in terms of formal parameters each of which can be determined by empirical investigation.

Although we are obviously still far from achieving such an integrated model of intonation, all of the models which I have somewhat disparagingly dubbed 'second generation' possess at least some of the desirable characteristics. Just as the accumulation of first-generation descriptions made it possible to develop such models in the late '70s, so the existence of such multi-language models will be the basis on which it will be possible to build a coherent general theory of intonation. The task for the '90s could thus be a concentration of collaborative (and/or competitive) work leading to the development of such a general theory.

REFERENCES

[1] BAILLY, G; BACH, M. LABOISSIERE, R.& OLESON, M. (1990) 'Generation of articulatory trajectories using sequential networks' *Proc. ESCA Workshop on Speech Synthesis* (Autrans), 67-70

[2] BRUCE, G. (1988) "How floating is focal accent ?" *Nordic Prosody* 4, 41-49

[3] CHAMBON, T. (1991) "Phonological interpretation of Fo variations in a Bantu language : Kinyarwanda".*Proc. Phon.Sci. XII* (vol. 2), 218-221

[4] CRYSTAL, D. (1975) 'Relative and absolute in intonation analysis' in *The English Tone of Voice* (Arnold, London), 74-83

[5] DI CRISTO, A. (in press) "Intonation in French". in Hirst & Di Cristo (eds) (in press).

[6] FUJISAKI, H. (1981), "Dynamic characteristics of voice fundamental frequency", *Proc. 8th FASE* (Venice)

[7] GOLDSMITH, J. (1976), *Autosegmental Phonology*, (Ph.D.thesis; MIT.)

[8] GOLDSMITH, J. (1990) *Autosegmental and Metrical Phonology.* (Blackwell, Oxford)

[9] GÅRDING, E. (1977) "The importance of turning points for the pitch patterns of Swedish accents.", in Hyman ed. (1977) *Studies in Stress and Accent.*, 27-35

[10] GÅRDING, E. (1983), "A generative model of intonation", in Cutler & Ladd (1983) *Prosody, Models and Measurements*, (Springer), 11-25

[11] GÅRDING, E. (in press), "Intonation in Swedish", in Hirst & Di Cristo (eds) (in press)

[12] 'T HART, J. & COHEN, A. (1973), "Intonation by rule : a perceptual quest.", *Journal of Phonetics* 1, 309-327

[13] 'T HART, J. COLLIER, R & COHEN, A.. (1990) *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*, (Cambridge University Press, Cambridge)

[14] HIRST, D.J. (1983) "Structures and categories in prosodic representations." in Cutler & Ladd (1983) *Prosody: Models and Measurements* (Springer, Berlin) 93-109

[15] HIRST, D.J. (1987) *La description linguistique des systèmes prosodiques : une approche cognitive.* (Thèse de Doctorat d'Etat, Université de Provence.)

[16] HIRST, D.J. (in press) "Intonation in British English" in Hirst & Di Cristo (eds.) (in press).

[17] HIRST, D.J. & DI CRISTO, A. (in press) "A survey of intonation systems." in Hirst & Di Cristo (eds). (in press)

[18] HIRST, D.J. & DI CRISTO, A. eds. (in press) *Intonation Systems : a Survey of Twenty Languages.* (Cambridge University Press; Cambridge)

[19] HIRST, D.J. NICOLAS, P. & ESPESSER, R. (1991) "Coding the F0 of a continuous text in French: an experimental approach." in *Proc. Phon.Sci. XII* (vol. 5), 234-237

[20] HOUSE, D. (1990) *Tonal Perception in Speech* (Lund University Press, Lund)

[21] LADD, D.R. (1986) "Intonational phrasing: the case for recursive prosodic structure." *Phonology Yearbook* 3, 311-340

[22] LIBERMAN, M & PIERREHUMBERT, J (1984) "Intonational invariance under changes in pitch range and length." in Aronoff & Oehrle (1984) *Language Sound Structure* 157-253

[23] LINDBLOM, B. LUBKER, J. & GAY, T. (1979) "Formant frequencies of some fixed-mandible vowels and a model of speech-motor programming by predictive simulation." *Journal of Phonetics* 7, 141-161.

[24] MEHLER, J. & DUPOUX, E (1990) *Naître Humain* (Seuil, Paris)

[25] OHALA, J.J. (199O) "There is no interface between phonology and phonetics : a personal view." *Journal of Phonetics*, 18, 153-171

[26] ÖHMAN, SVEN & LINDQVIST, J. (1966), "Analysis by synthesis of prosodic pitch contours.", *Proc.18th Int. Congress. of Psychology* (Moscow)

[27] PIERREHUMBERT, J.. & BECKMAN, M.. (1988), *Japanese Tone Structure.*, (MIT Press; Cambridge, Mass.)

[28] PIERREHUMBERT, J. (1980) *The Phonology and Phonetics of English Intonation.* PhD thesis; MIT.

[29] SELKIRK, E.(1984) *Phonology and Syntax : the Relation between Sound and Structure.* (MIT Press; Cambridge, Mass.)