Speech prosody: from acoustics to interpretation.

Daniel J. Hirst

Dear Professor Wu Zongji,

It is a tremendous pleasure and honour to be able to be here today to wish you a *Happy Birthday* and *Many Happy Returns!*

I am very proud to be here among so many eminent scholars who have travelled from all over the world to greet you today.

Many of them are far more qualified than I am to talk about your career and your contribution to the speech sciences.

Instead, let me offer you this modest poem:

In the shade of the iron pagoda, Xiao Wu,
you wonder at destiny, the gift
of names, the weight of words, the way
to save our ancestors. The trees,
dancing in the wind, whisper
"lift up our load and carry it
onward as far as you can".
Lao Wu,
yours is a long journey, on your way
you have added many more ripe fruit
to those you have brought from afar!
Wu Lao,
may we walk some way with you along your path?

Daniel Hirst

Speech prosody: from acoustics to interpretation

Daniel Hirst *

Abstract

The way in which prosody contributes to meaning is still, today, a poorly understood process corresponding to a mapping between two levels of representation for neither of which there is any general consensus. It is argued that annotation of prosody generally consists in describing both prosodic function and prosodic form, but that it would be preferable to clearly distinguish the two levels. One elementary annotation system for prosodic function: IF-annotation, is, it is argued, sufficient to capture at least those aspects of prosodic function which influence syntactic interpretation. The annotation of prosodic form can be carried out automatically by means of an F0 modelling algorithm, MOMEL, and an automatic coding scheme, INTSINT. The resulting annotation is underdetermined by the IF-annotation, but defining mapping rules between representations of function and representation of form could provide an interesting means of establishing an enriched functional annotation system through analysis by synthesis.

1. Introduction

Everybody agrees that prosody contributes to the meaning of an utterance. In fact, when there is a discrepancy between the prosody of the utterance and its overt semantic content we usually trust the prosody rather the semantics. So when someone says:

It's so exciting!

with a bored tone of voice, we tend to believe he is bored even though he has overtly said the opposite. Similarly if the sentence:

He's got nice handwriting…!

is pronounced with a falling-rising nucleus on 'handwriting', the utterance is quite likely to be interpreted as a criticism even though it is overtly a compliment.

For a 'real life' example of the complex way in which prosody can contribute to the interpretation of an utterance, the following sentence French sentence was pronounced by a newsreader on the French national radio France Inter[1]:

> *Il semble que les policiers sont à deux doigts d'arreter Spaggiari, mais il faudra qu'il fassent vite pour trouver la cachette de l'ancien parachutiste.*
> (It seems that the police are on the point of arresting Spaggiari but they'll have to act quickly to find the hiding place of the former parachutist)

Reading this sentence, the first interpretation which springs to mind is that the police are looking for two escaped prisoners, one named Spaggiari and the other a former parachutist. The intonation used by the speaker, however, with a falling pitch on 'vite' and a low flat pitch on 'pour trouver la cachette de l'ancien parachutiste', made it clear that the former parachutist in question was Spaggiari himself rather than being someone else who had escaped from prison in his company. Similar interpretations could be obtained by different readings of the English translation.

---

[1] France Inter, Informations 13-14. 12 mars 1977

There is, today, no general consensus on the way in which the prosody of an utterance contributes to its meaning. The last six chapters of Couper-Kuhlen (1986) constituting about half of the book, give a well-documented account of the various ways in which the problem of the meaning of intonation has been approached in the literature. The fact that intonation meaning can be approached in so many different ways, for results which, as Couper-Kuhlen (p209) admits:

"are rather modest indeed"

seems to indicate that we have still not got properly started on the analysis of intonational meaning. As Cruttenden(1986:184) puts it:

"it is not yet even clear what sorts of meanings are involved."

Surprisingly, in the literature on prosody, there are fewer publications addressing this subject as compared to phonological, phonetic or acoustic analyses for example, despite the fact that nearly everyone agrees that this is the central question in the field.

There are a number of possible explanations for this at first sight rather surprising fact. The simplest one is that researchers tend to be specialists in one specific domain and that the interaction between phonology and interpretation, as implied in the title of this paper, requires a knowledge of two fields, phonology on the one hand and syntax, semantics and pragmatics on the other. Few people working in the field of syntax, semantics or pragmatics have detailed knowledge about prosody and the reverse is just as true.

A further explanation comes from the fact that phonological representation and syntactic/semantic/pragmatic interpretation tend to be very theory dependent. To convince an audience that your explanation of the way in which intonation contributes to meaning is correct, you have to be able to convince them that both your phonology and your interpretation are right.

Not only there is no consensus on the explanation for the way that prosody contributes to meaning, there is not even a real consensus on the way that prosody should be represented phonologically, nor on the way in which we should represent the sort of meanings that prosody contributes.

In the rest of this paper I look at the question of prosodic annotation and the distinction between prosodic function and prosodic form. I conclude that a clear separation between these two would be highly desirable and could lead to new insights into the way in which a phonological representation of prosody can be mapped onto a functional representation

2. Prosodic annotation.

2.1 The ToBI system a standard for prosodic annotation.

In an attempt to meet this need for consensus in prosodic representation, a group of linguists and engineers, mostly American, came to an agreement over a system of representation for the prosody of American English (Silverman et al. 1992). This system, which they called **ToBI**, an acronym for Tones and Break Indices, proposed to represent the prosody of an utterance by means of an alphabet of discrete symbols representing the different pitch accents which had been described in American English, decomposed into sequences of symbols **H** and **L** (for high and low tones respectively), together with a scalar representation of the degree of separation between consecutive words, going from 0 (absence of break) to 4 (major intonation unit break). Within each pitch accent, one tone symbol is accompanied by the diacritic [*] indicating that the tone in question is directly associated with the accented syllable of the word. Besides the break indices, boundaries are also marked by the presence of a "boundary tone", again either **H** or **L**. These are distinguished from the tones belonging to the pitch

accents by the presence of a diacritic symbol: [**-**] for the so-called "phrase accent" (in fact a phrase boundary tone) and [**%**] for major intonation boundaries.

This system was proposed, and accepted, as a standard for the description of the prosody of American English utterances and rapidly became the most widely used prosodic annotation system in the world. Although the system was originally designed uniquely for the description of American English, it has rapidly been adapted for a number of other dialects and languages including German, Italian, Japanese and Chinese. It is worth noting, however, that the authors of the system themselves warn against using ToBI indiscriminantly to describe other languages.

One of the principal authors of ToBI, Janet Pierrehumbert, insists on the fact that ToBI was based on a detailed analysis and a particular theory of the intonation system of American English. To describe the prosody of a language using this system it is indispensable to begin with an exhaustive inventory and phonological analysis of the possible pitch accents and boundary tones of the language.

Despite these warnings, a number of publications have used an adaptation of ToBI to describe the pitch patterns of languages for which there is not yet a complete phonological description of the intonation system. In doing so they attempt to use ToBI as the prosodic equivalent of the International Phonetic Alphabet.

The authors of ToBI (ToBI 1999) clearly state (on the official ToBI website, that:

> *Note: ToBI is not an International Phonetic Alphabet for prosody. Because intonation and prosodic organization differ from language to language, and often from dialect to dialect within a language, there are many different ToBI systems, each one specific to a language variety and the community of researchers working on that language variety.*

2.2 ToBI or not ToBI

Ten years after the ToBI standard for prosodic annotation of American English was first proposed (Silverman et al. 1992), one of the co-authors of the original paper, Colin Wightman (1992), presented a critical evaluation of the usefulness for speech technology of this annotation system.

One of the major aims of the ToBI project was to provide a system which would have a high level of inter-transcriber agreement. Wightman notes that, while considerable cross-transcriber agreement has indeed been found for the identification of prominences and boundaries, the agreement is far less consistent for the type of pitch accent or the type of boundary. He claims furthermore that much of the motivation for large-scale manual transcription of speech corpora is today obviated by the general availability of software capable of extracting prosodic information automatically from the speech signal as well as by the massive increase in size of computer memory which makes it possible today to carry out on personal computers analyses of large corpora, which ten years ago could only have been performed by mainframe computers in specialised institutes.

Wightman's conclusion is that because of its labour-intensive cost, manual transcription should be reserved for those aspects of prosody which untrained listeners actually hear: he formulates this as the maxim:

"Transcribe what you hear!"

This maxim could be interpreted in a number of ways. Listeners obviously hear many different things, including some of which they are not consciously aware. It has been shown, for example, (Hawkins and Slater 1994., Heid and Hawkins, 1999) that listeners may be generally incapable, under normal listening conditions, of distinguishing utterances produced by speech synthesis implementing very fine differences in the way

in which co-articulation phenomena are handled. When the same utterances are heard in adverse conditions, such as with heavy background noise, however, there is a considerable difference in performance on intelligibility tests between the two sets of utterances. Very fine details of acoustic information, then, can be seen to contribute to the robustness of speech perception and comprehension. It seems difficult to claim that listeners do not hear these differences, even though they may be unaware of doing so.

One way of interpreting Wightman's maxim would be to say that manual transcription should be reserved for those aspects of prosody that contribute to the listener's interpretation of the utterance. Transcribers, in other words, should be attentive to prosodic function rather than to prosodic form. In this way, the transcriber is required to perform a task of linguistic interpretation rather than a meta-linguistic task of phonetic analysis. As is well known in psycholinguistic studies, meta-linguistic tasks performed by untrained subjects entail considerable problems of interpretation.

In this paper, I suggest that a systematic distinction between function and form is a highly desirable aspect of a prosodic annotation system. I outline some specific proposals in this area and suggest that such a multi-level system of annotation could be of interest for speech synthesis and automatic speech recognition as well as for fundamental research into the linguistic analysis of speech prosody.

3. Function and Form in prosody

Like all linguistic phenomena, prosody has both function and form. In many systems of prosodic annotation, perhaps most, the two levels of representation are intimately intertwined. There are, however, a number of reasons for distinguishing these levels.

First of all, many prosodic functions seem to be quasi-universal (Hirst & Di Cristo 1998), in nearly all languages prosody contributes in some way to lexical identity (via

tone, quantity and accent)[2], expressing prominence, boundaries, non-finality etc. In the same way, prosodic forms are certainly universal all languages use rising and falling pitch, longer and shorter segments, etc. The mapping between form and function, however, is certainly not universal. If it were so, we should expect all languages to use the same prosodic forms to express the same meanings and this is clearly not the case.

A second reason is that studying the relationship between prosodic form and function becomes rather circular if a clear distinction between the two levels is not made. To take a fairly trivial example, if we were to make use of a prosodic annotation system that distinguished two rising intonation contours, calling one a continuation rise and the other an interrogative rise, there would be little point in examining the correlation between the distribution of the two patterns with respect to syntactic or pragmatic criteria unless the two patterns could be distinguished entirely on the basis of their formal characteristics. Halliday (1967:21), for example describes the difference between continuation rises and interrogative rises as follows:

> *The difference, though gradual, is best regarded as phonetic overlap (...) the one being merely lower than the other (...) But the meanings are fairly distinct. In most cases the speaker is clearly using one or the other; but sometimes one meets an instance which could be either.*

Halliday is clearly basing the distinction between the two types of rises on prosodic function although he presents them as if they were distinct prosodic forms.

Such confusion of prosodic form and prosodic function is far more widespread than is commonly realised and this mixing of levels can only be prejudicial to linguistic analysis. The use of such hybrid annotation is not restricted to any particular school or tradition of prosodic analysis.

Annotation systems developed in the British school, for example (such as O'Connor & Arnold 1961, Hallliday 1967, Crystal 1969, Cruttenden 1986, Couper-Kuhlen 1986),

[2] French in this respect is a rather exceptional language in that lexical representations need to include neither tone, accent nor quantity.

used different symbols to annotate similar types of pitch movement (rising, falling, falling-rising etc) depending on the prosodic function of the pitch movement considered as a pre-nuclear accent or a nucleus.

The ToBI annotation system also combines representations of prosodic form (H, L) with representations of prosodic function (- * %).

4. Representing prosodic function

I mentioned above that inter-transcriber agreement is in general far higher when transcribers are asked to concentrate on prosodic function rather than on prosodic form. This suggests that a first approximation for an annotation scheme for prosodic function could be to adapt ToBI by dropping the tonal specification and keeping only the boundaries and prominences. For an adaptation of this type cf. Wightman & al. (2000). We can call this toneless ToBi or StarBI annotation.

A number of years ago, I proposed a slightly more elaborate functional annotation system (Hirst 1977) with a system taking into account four degrees of prominence (unstressed, stressed nuclear and emphatic) and two types of prosodic boundary terminal and non-terminal. An example of this type of annotation is the following which includes all the symbols: ' [stress], ° [nucleus], | [boundary], + [non-terminal boundary], || [terminal boundary] __ [emphasis]:

| If you 'can't °lift it + 'ask °<u>Peter</u> to 'help you ||

Corresponding to an utterance like the following:

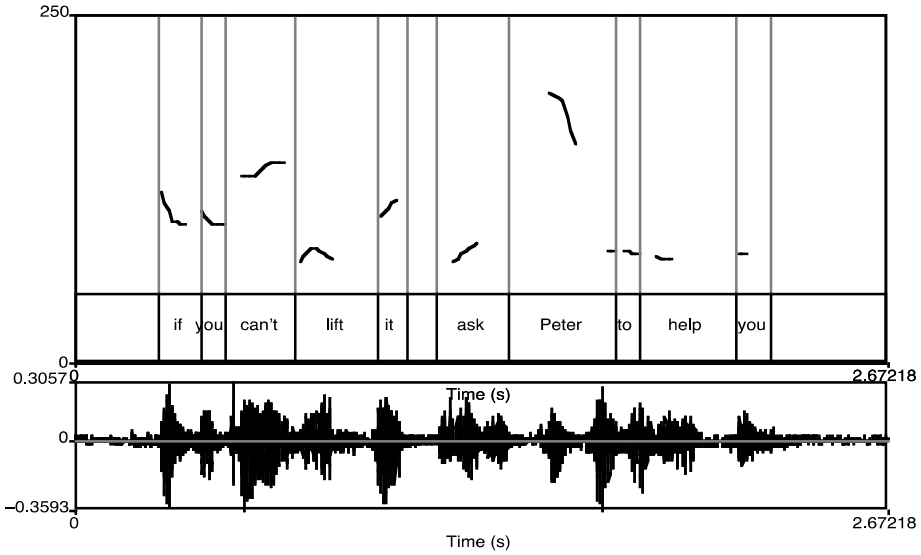
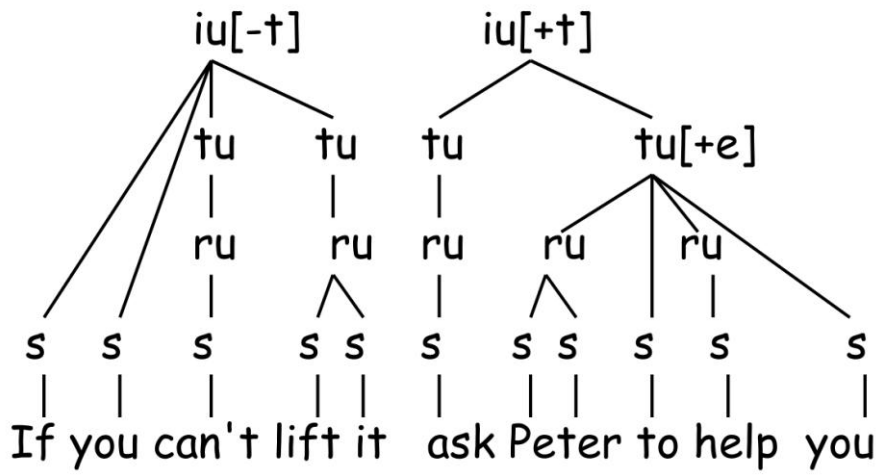


Figure 1. Signal and F0 of the utterance "If you can't lift it, ask Peter to help you."

I refer to this system, to which I return below, as IF annotation (which can be glossed as either Intonative Features, the title of the 1977 book or alternatively as Intonation Functions). In Hirst (op cit), I argue that this annotation is sufficient to account for all those aspects of prosodic representation which contribute directly or indirectly to syntactic interpretation.

The original formulation of the IF annotation system was presented as a set of distinctive features. Within the framework of non-linear phonology, assuming that segments are grouped into higher level phonological constituents, specifically into syllables, rhythm units[3], tonal units and intonation units, only the features [± emphatic] and [± terminal] need to be retained as features. The utterance above could then be transcribed:

[3] For arguments in favour of the tonal unit as a phonological constituent for English and French cf Hirst 1988, for the rhythm unit cf. Jassem 1952, Bouzon & Hirst 2003.



which is formally equivalent to the linear IF transcription.

5. Representing prosodic form.

As Wightman (2002) observed, the need for manual annotation of prosodic form is far less obvious today with the widespread availability of automatic algorithms for pitch extraction and stylisation. My colleagues and I have suggested (Hirst & al 2000) that the mapping between prosodic form and prosodic function should involve a number of different levels of representation including a level of phonetic representation, consisting of scalar values directly related to the acoustic signal, and a level of surface phonology, which unlike the phonetic representation, codes the prosodic form as a sequence of discrete symbols but which are still directly related to the acoustic signal. We also propose a more abstract underlying phonological representation of prosodic form, which we assume is directly related to the representation of prosodic function. In the next sections I outline briefly the nature of these different levels of representation.

5.1 Phonetic representation

The MOMEL algorithm developped in Aix-en-Provence (Hirst & Espesser 1993), provides an automatic phonetic representation of a fundamental frequency curve. The algorithm is often referred to as a stylisation of fundamental frequency but it should

more properly be called a model since it consists in factoring the raw fundamental frequency curve into two components without any loss of information. These are a macroprosodic component, consisting of a continuous smooth curve (represented as a quadratic spline function) corresponding to the linguistic function of the contour, and a microprosodic component consisting of deviations from the macroprosodic curve caused by the nature of the phonematic segment (voiced/unvoiced obstruent, sonorant, vowel etc) (cf Di Cristo & Hirst 1986). The output of the algorithm is a sequence of target points which are sufficient to define the macroprosodic component of the fundamental frequency when used as input to a quadratic spline function.

Momel is currently available in a number of different implementations in various speech-analysis environments including Mes for Unix (Espesser 1996), SFS for Windows (Huckvale 2000) as well as in the multi-platform system Praat (Boersma & Weenink 1995-2003)4.

A recent evaluation of the algorithm (Campione 2000) was carried out using recordings of the continuous passages of the Eurom1 corpus for five languages (English, German, Spanish, French, Italian) in all a total of 5 hours of speech). The evaluation estimated a global precision of 97.6% by comparison with manually corrected target estimation. Compared to the 46982 target points provided by the automatic analysis, 3179 were added manually by the correctors and 1107 removed. The algorithm gave only slightly worse results (93.4% precision) when applied to a corpus of spontaneous spoken French. The majority of these corrections involved systematic errors, in particular before pauses, which an improvement of the algorithm should eliminate.

The output of the algorithm as a sequence of target points is particularly suitable for interpretation as a sequence of tonal segments such as the INTSINT representation described below, but the theory neutral nature of the modelling, together with its

reversibility, has allowed the algorithm to be used as input for other types of annotation including ToBI (Wightman & Campbell 1995, Maghbouleh 1998) and the Fujisaki model (Mixdorff 1999).

5.2 Surface phonological representation

The prosodic annotation system INTSINT was based on the descriptions of the surface patterns of the intonation of twenty languages (Hirst & Di Cristo eds 1998) and was used in that volume for the description of nine languages (British English, Spanish, European Portuguese, Brazilian Portuguese, French, Romanian, Bulgarian, Moroccan Arabic and Japanese).

Intonation patterns are analysed as consisting of a sequence of tonal segments, defined in one of two ways: either globally with respect to the speaker's pitch range (**Top**, **Mid** or **Bottom**) or locally with respect to the preceding target (**Higher**, **Same** or **Lower**) with an iterative variant of these locally defined targets (**Upstepped**, **Downstepped**) assuming that an iterative tone can be followed by the same tone whereas a non-iterative tone cannot and furthermore that the iterative tones correspond to a smaller pitch interval than the non-iterative ones.

This transcription system, originally designed as a tool for linguists transcribing the intonation of utterances of different languages, was intended to provide at least a first approximation to a prosodic equivalent of the International Phonetic Alphabet. As we saw above this is specifically not the case for the ToBI system.

In the case of INTSINT, it was intended from the first that the transcription should be convertible to and from a sequence of target points. A first version of an algorithm for converting between Momel and INTSINT was described in (Hirst & al., 2000) . An extension of the system to annotate duration and timing has also been proposed (Hirst 1999).

A simpler and more robust algorithm has since been developed (Hirst 2000)5. In this version, target points are coded on the basis of two speaker dependent parameters: key and range. Given these two parameters, the absolute tones are defined as the limits of the speaker's pitch range (**Top** and **Bottom**) assumed to be symmetrical around the central value (**Mid**). The relative tones are then defined by an interval between the preceding target point ($\mathbf{P_{i-1}}$) and the two extreme values taken as an asymptote for these targets as in the following:

$$P_i = P_{i-1} + c.(A-P_i)$$

where **A** is either **T**, (for **H** and **U**) or **B** (for **L** and **D**) and where **c** is set at 0.5 for the non-iterative targets **H** and **L** and at 0.25 for the iterative targets **U** and **D**.

This algorithm, applied to the targets of the French and English passages of the Eurom1 corpus (Chan & al 1995), was optimised over the parameter space :

key = mean ± 50 (in Hz)
range [0.5, 2.5] (in octaves).

Interestingly, the mean optimal range parameter resulting from this analysis was not significantly different from 1.0 octave. It remains to be seen, however, how far this result is due to the nature of the EUROM1 corpus which was analysed (40 passages consisting each of 5 semantically connected sentences) and whether it can be generalised to other speech styles and other (particularly non-European) languages.

The symbolic coding of the F0 target points obviously entails some loss of information with respect to the original data, unlike the Momel analysis which is entirely reversible. The loss of information is, however, rather small as can be seen from Figure 2 which illustrates the output from the optimised INTSINT coding compared to the original target points for a complete five sentence passage from the Eurom1 corpus.

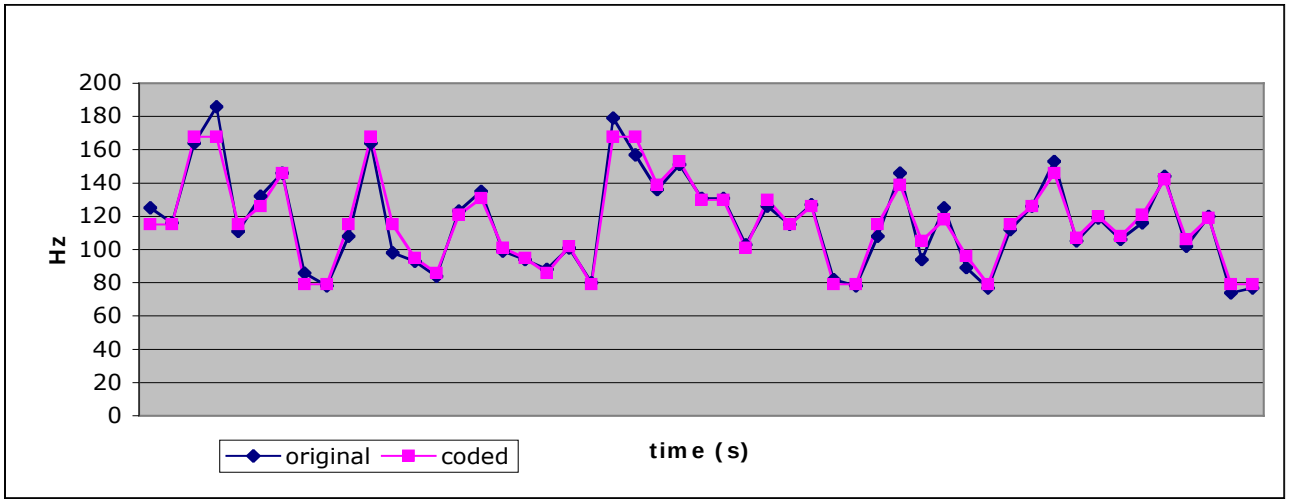


Figure 2. Coding of the F0 targets from a sample passage from the Eurom1 corpus showing the original target points estimated by the Momel algorithm and the target points derived from the optimised INTSINT coding.

6. Deriving prosodic Form from prosodic Function

As I mentioned above, Momel and INTSINT provide reversible representations of intonation patterns since not only can they be derived automatically from the acoustic signal but it is also possible to convert a sequence of INTSINT symbols, together with two speaker/utterance dependent parameters key and range, into a sequence of target points which can then be converted to a smoothed fundamental frequency curve.

The example given above:

| If you 'can't °lift it + 'ask °Peter to 'help you ||

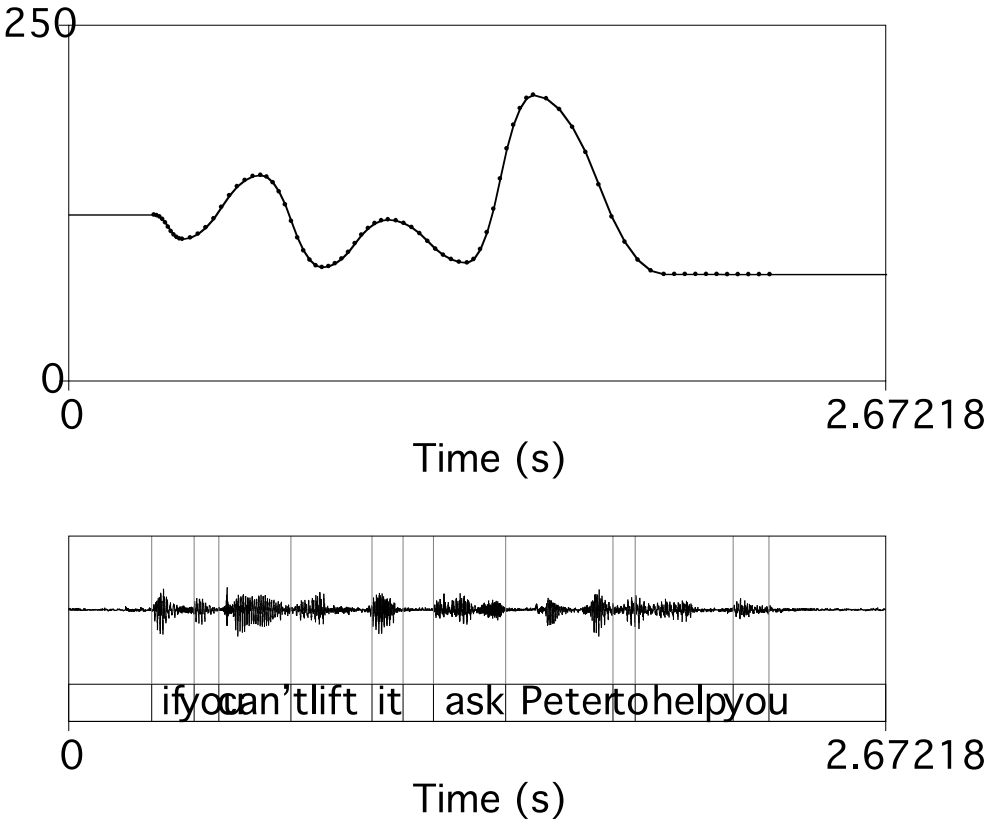
would be converted to :

If you can't lift it ask Peter to help you.
M H B H M B T B B

which, in turn, can be converted to an appropriate sequence of tonal targets as input to a speech synthesis system such as:

If you can't lift it ask Peter to help you.
135 163 95 143 135 95 191 95 95

which can then be used to produce a continuous F0 curve:



Work in progress involves the elaboration of mapping rules between IF annotation and INTSINT annotation for both English and French. Converting IF to INTSINT is fairly simple. An implementation for French, is described in (Di Cristo & al 2000). The inverse mapping, however, is not currently feasible, since IF in its present state can generate only a subset of possible and observed INTSINT patterns.

Thus for example [+emphatic] in British English, will generally correspond to a high falling nuclear pitch accent when followed by a [+terminal] boundary but to a rising-falling nuclear pitch accent when followed by a non-terminal boundary. There are, however, a number of secondary characteristics which often, but not always, accompany emphatic nuclear pitch accents. The high falling pattern, for example, is often preceded by an upstepping head. This corresponds to the global pattern which has sometimes been called the "surprise/redundancy" pattern. The effect of the upstepping head is to reinforce the fact that the final fall is higher than the preceding accent. This

characteristic, while very common for emphatic terminal pitch patterns, is by no means the only possibility and seems to represent a separate choice on the part of the speaker.

The fact that an upstepping head begins with a low accent may furthermore be reinforced by a high onset for any preceding unstressed syllables (high pre-head). Once again, while this is a common characteristic of upstepping heads it is by no means necessary and it is not restricted to this context, either.

Similarly, a falling-rising nuclear pattern (emphatic non-terminal) is frequently preceded, in British English, by a sequence of falling pitch patterns on the pre-nuclear accents (the head). This is in fact the only context, in British English, (contrary to American and Scottish English) where this type of pattern is fairly systematic although it is by no means impossible in other contexts.

In (Hirst 1998) I argue that surface phonological representations for non-emphatic and emphatic intonation patterns in British English can be derived from rather abstract underlying phonological representations which are quite naturally related to the prosodic structure represented in the IF annotation.

In this approach, two prosodically very different languages such as English and French can be characterised by means of a small number of abstract prosodic parameters. French, like other Romance languages would be characterised as having a right-headed Tonal Unit while English, like other Germanic languages would have a left-headed Tonal Unit. The underlying tonal template for French, in this analysis, would be the sequence [L H] whereas in English the underlying sequence would be [H L]. It seems furthermore that the hierarchy of prosodic constituents is different in English and French with English possessing a rhythm unit which is at a lower level than the tonal unit (as suggested by Jassem (1952) over 50 years ago) whereas in French it seems more appropriate to consider the tonal unit as being on a lower level than the rhythm

unit. One of the results of this parametrisation of the phonology of prosodic systems is that for French, unlike for English, there is no distinctivity for the "nuclear" pitch accent since the possibility of the nuclear accent in English occuring on a non-final stress (without emphasis) is a consequence of the possibility of grouping several rhythm units into a single tonal unit, which is not possible in French apart from in emphatic patterns.

7. Deriving prosodic function from prosodic form

The ultimate aim of describing the prosody of natural language utterances is to provide a deeper understanding of the way in which prosody contributes to the interpretation of these utterances. Such a goal clearly has implications for speech technology since, as Wightman (op cit) notes, despite the considerable research invested in the transcription of prosody of various different languages in the last decade, the actual implementation of prosody in TTS or ASR applications is remarkably limited. Paradoxically, as Ostendorf has noted (2000), speech technology is even more in need of prosodic aids than human speakers, in both production and perception, since computers have far less knowledge of the world than humans to help them to interpret utterances.

The under-determination of the INTSINT representation with respect to IF annotation, suggests a strategy of analysis by synthesis which seems rather promising.

In this approach, a preliminary IF annotation is used to generate an INTSINT representation, which is then compared to the annotation derived from the actual recording. Systematic differences can then be used to either correct the mapping rules or to extend the IF annotation system which in its present state is obviously rather rudimentary. In the emphatic examples which we discussed above, for example, we might decide that the high falling nuclear pattern, the upstepping head and the high pre-head constitute three independent choices for the speaker with respect to the emphatic nature of the utterance.

This in turn suggests a number of experimental paradigms to examine the orthogonality of such subsets of intonation patterns which we intend to explore in more detail in future work.

References

Auran, C. 2003. Momel and Intsint package. http://www.lpl.univ-aix.fr/~auran/

Boersma, P., & Weenink, D. 1995-2003. Praat: a system for doing phonetics by computer. htttp://www.fon.hum.uva.nl/praat/

Bouzon, C & D.J.Hirst 2004. Isochrony and prosodic structure in British English. in Bel & Marlien (eds) *Proceedings of Speech Prosody 2004.*

Campione, E. 2001. *Etiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie*. Doctoral thesis.. Aix-en-Provence: Université de Provence..

Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinas, G., Kvale, L., Lamel, L., Lindberg, L., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., & Zeiliger, J. 1995. EUROM: a spoken language resource for the EU. *Proceedings of the 4th European Conference on Speech Communication and Speech Tecnology, Eurospeech '95*, (Madrid) 1, 867-880.

Couper-Kuhlen, E. 1986. *An Introduction to English Prosody.* London: Arnold.

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

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

Di Cristo, A. & Hirst, D.J. 1986. Modelling French micromelody. Analysis and synthesis. Phonetica,. 43, 11-30.

Di Cristo, A.; Di Cristo, P. & Véronis, J. 1997. A metrical model of rhythm and intonation for French text-t-speech synthesis. in A. Botinis (ed.) Intonation: Theory, Models and Applications (ESCA), 83-86.

Espesser, R. 1996. MES : Un environnement de traitement du signal. *Proceedings XXIe Journées d'Etude sur la Parole* 1996 June 10-14 : Avignon, France, p. 447..

Halliday, M.A.K 1967. *Intonation and Grammar in British English*. The Hague: Mouton.

Hawkins, S., and Slater, A. 1994. Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *ICSLP 94 (Proceedings of the 1994 International Conference on spoken Language Processing)* 1, 57-60.

Heid, S. and Hawkins, S. 1999. Synthesizing systematic variation at boundaries between vowels and obstruents. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A.C. Bailey (eds.), *Proceedings of the XIVth International Congress of Phonetic Sciences*. University of California, Berkeley, CA. 1, 511-514.

Hirst, D.J. 1977. *Intonative features. A syntactic approach to English intonation*. (= Janua Linguarum series minor 139), Mouton, The Hague.

Hirst, D.J. 1988. Tonal units as phonological constituents: the evidence from French and English intonation. in H. Van der Hulst & N. Smith (eds) 1988, *Autosegmental studies in pitch accent*. Foris, Dordrecht, 151-165.

Hirst, D.J. 1998. Intonation in British English. in Hirst & Di Cristo (eds.) 1998.

Hirst, D.J., 1999. The symbolic coding of segmental duration and tonal alignment. An extension to the INTSINT system. *Proceedings ICSLP '99*.

Hirst, D.J. 2000. Optimising the INTSINT coding of F0 targets for multi-lingual speech synthesis. *ISCA Workshop: Prosody 2000 Speech recognition and Synthesis* (Kraków, October 2000).

Hirst, D.J. & Di Cristo, A. 1998. A survey of intonation systems. in Hirst & Di Cristo (eds) 1998, .

Hirst, D.J. & Di Cristo, A. (eds). 1998. *Intonation systems. A survey of twenty languages*. Cambridge University Press. Cambridge.

Hirst, D.J., Di Cristo, A. & Espesser, R. 2000. Levels of representation and levels of analysis for the description of intonation systems. in M. Horne (ed) 2000, 51-87.

Hirst, D.J. & Espesser, R. 1993. Automatic modelling of fundamental frequency curves using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15, 71-85.

Horne, M. (ed) *Prosody : Theory and Experiment. Studies Presented to Gösta Bruce*. Dordrecht: Kluwer Academic Publishers.

Huckvale, M. 2000. Speech Filing System. Tools for speech research .
http://www.phon.ucl.ac.uk/resource/sfs/

W. Jassem. 1952. *Intonation of Conversational English (Educated Southern British)*. Wroclaw, Wroclawskie Towarzystwo Naukow.

Maghbouleh, A. 1998. ToBI accent type recognition. *Proceedings ICSLP '98.*

Mixdorff, H. 1999. A novel approach to the fully automatic extraction of Fujisaki model parameters. *ICASSP 1999.*

OConnor, J.D. and Arnold, G.F. 1961. *Intonation of Colloquial English*. London: Longman (2nd edition, 1973).

Ostendorf, M. 2000. Prosodic boundary detection. in M. Horne (ed) 2000, 263-279.

Rolland, G. 2000. Automatic stylisation of fundamental frequency with MOMEL.
http://www.icp.inpg.fr/~rolland/my_work/index.html

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. ToBI: a standard for labelling English prosody. Proceedings ICSLP92, 2, 867- 870, Banff, Canada.

TobI 1999. http://www.ling.ohio-state.edu/~tobi/

Wightman, C. & Campbell, N. 1995. Improved labeling of prosodic structure. *IEEE Trans. on Speech and Audio Processing.*

Wightman, C. 2002. ToBI or not ToBI? in B.Bel & I.Marlien (eds) *Proceedings of Speech Prosody 2002*, Aix en Provence.

Wightman, C. W., Syrdal, A. K. et al., 2000. Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative speech synthesis, in *Proceedings ICSLP*, vol. 2, pp. 7174.