Form and function in the representation of speech prosody

Daniel J. Hirst

Form and function in the representation of speech prosody.

Daniel Hirst

CNRS, UMR 6057 Parole et Langage, Université de Provence
Aix-en-Provence, France
email: *daniel.hirst@lpl.univ-aix.fr*

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 has been 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.

Keywords:
speech prosody, annotation, form, function

Corresponding author: Daniel Hirst, CNRS UMR 657, Laboratoire Parole et Langage, Université de Provence, 29 avenue Schuman, 13621 Aix en Provence, France
telephone: +33 4-4295-3628, fax: +33 4-4259-5096

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:

(1) 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:

(2) He's got nice handwriting…!

is pronounced with a falling-rising pitch 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 was pronounced by a newsreader on the French national radio *France Inter*[1]:

(3) Il semble que les policiers sont à deux doigts d'arrêter Spaggiari, mais il faudra qu'ils 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)

If we read this sentence, without hearing it spoken, one likely interpretation is that the police are looking for two escaped prisoners, one of whom is named Spaggiari and the other who is 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

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

who had escaped from prison in his company. Similar interpretations can be obtained by different readings of the English translation.

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, requires a knowledge of widely different areas of research, 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 of course just as true.

A further explanation comes from the fact that phonological representation and syntactic/semantic/pragmatic interpretation tend to be very theory dependent. 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. To convince others 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.

In the rest of this paper, as a potential solution to this problem, I look at the question of prosodic annotation and the distinction between prosodic function and prosodic form. I argue that a clear separation between the two is 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 and in which a more elaborate functional representation might be established empirically by means of analysis by synthesis.

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 a syllable carrying a pitch accent. 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 originally designed for the description of one specific dialect, it has rapidly been adapted for a number of other dialects and languages including German, Italian, Japanese, Korean and Chinese (Jun (ed) 2005).

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 (2000), 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 where 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, however, clearly state on the official ToBI website (ToBI 1999), 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 (2002), 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, however, 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, as summarised in Table 1.

INSERT TABLE 1 ABOUT HERE

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 formal 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 departments.

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 actually 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, however, such as with heavy background noise, 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 (in the sense that their auditory systems fail to register them) even though they may be unaware of doing so.

One way of reformulating Wightman's maxim would be to say that manual transcription should be reserved for those aspects of prosody that refer directly to the listener's **interpretation** of the utterance. The fact that a given syllable is interpreted as being more prominent than another or as being followed by a prosodic boundary seems clearly to reflect a difference of meaning rather than a simple difference of form. The fact that a given syllable appears longer than another, or that it seems louder or is pronounced on a higher pitch is, by contrast, an analysis of what the utterance **sounds like** rather than what it **means**. Under this formulation, we could say, then, that transcribers 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 (cf for example Scarnà & Ellis 2002), 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 of speech prosody

Like all linguistic phenomena, speech prosody has both function and form. In many systems of prosodic annotation, perhaps most, the two levels of representation are intimately intertwined.

In nearly all languages, prosody contributes in some way to lexical identity (via tone, quantity and accent)[2], to expressing prominence, emphasis, boundaries, non-finality etc. (Hirst & Di Cristo 1998) as well as to a large number of still rather poorly understood means of expressing such things as dialogue structure, speech acts as well as general affect.

Prosodic forms are also certainly universal, in the sense that all languages make some use of contrasts between rising and falling pitch, between longer and shorter segments, etc.

There are, however, a number of reasons for wishing to distinguish form and function. First of all, although many prosodic functions and prosodic forms seem to be quasi-universal, the mapping between form and function, 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. The nearly universal tendency to associate a final lowering of pitch with statements and a final raising of pitch with questions, for example, is contradicted by a number of languages which do not systematically use a final rise for questions (e.g. Danish, Finnish and Western Arabic, Hirst & Di Cristo 1998), although in such languages, questions are often associated with a global raising of pitch or an incomplete final lowering. There are, furthermore, languages and dialects where a default declarative utterance is regularly produced with a final rise. These include certain varieties of Scandinavian languages as well as a number of urban dialects of Northern Britain (e.g. Newcastle, Glasgow, Belfast, Liverpool, Manchester…).

A further reason for clearly separating the two levels is that studying the relationship between prosodic form and function becomes rather circular if a clear distinction between them is not made.

To take a fairly simple 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 (clearly a functional distinction), 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

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

their formal characteristics. Halliday (1967: 21), 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 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, Halliday 1967, Crystal 1969, Cruttenden 1986, Couper-Kuhlen 1986), 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. For example a high-falling nuclear pitch accent on the word "no" is annotated: [`No!] whereas a high-falling pitch accent on a pre-nuclear accent on the same word in the expression “No way!” would be transcribed [↘No `way]. This means that the symbols are in fact combining two types of information: prosodic form (high falling pitch) and prosodic function (nuclear vs. pre-nuclear accent)

The ToBI annotation system as summarised above also combines representations of prosodic form (H, L) with representations of prosodic function (- * %) in so far as the latter symbols convey aspects of prosodic structure which are clearly expressions of what prosody does in language (i.e. its function) rather than what prosody sounds like (its form). A similar analysis could be applied to other descriptive frameworks derived from the original ToBI system (Ladd 1996, Grabe et al. 2001, Gussenhoven 2002, etc).

4 Representing prosodic function

I mentioned above that inter-transcriber agreement is, in general, far higher when transcribers are asked to concentrate on underlying prosodic structure rather than on prosodic form and I argued that this is because such underlying structure is an aspect of what prosody does in language, i.e. of prosodic function. 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 et al. (2000). We can call this rudimentary functional system *Toneless ToBi* or *StarBI* annotation.

A number of years ago, I proposed a slightly more elaborate functional annotation system (Hirst 1977), expressing 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 given in (4) which includes all the symbols used in the system: **stress**: ['], **nucleus** [°]; **boundary**: [|],**emphatic nucleus**: [!][3], **non-terminal boundary**: [+]; and **terminal boundary**: [‖]:

(4) | *If you 'can't °lift it + 'ask !Peter to 'help you* ||

This annotation corresponds to a pronunciation like that illustrated in Figure 1.

The terms **stress**, **boundary** and **emphasis** are perhaps familiar enough not to require further explanation although of course there are several non-trivial questions relating to each of these categories. The term **nucleus**, borrowed from the British tradition of intonation analysis (cf Cruttenden 1996 for example) refers to the principal (final or non-final) accent of an intonation unit. The distinction between **terminal** and **non-terminal** boundary refers to the listener's interpretation of the intonation unit as being finished or unfinished.

I refer to this system, to which I return below, as *IF annotation* (which can be glossed as either *Intonative Features*, the title of my 1977 book or alternatively

[3] The emphatic nucleus was represented by underlining the corresponding syllable in Hirst (1977). The [!] mark has the advantage that it is an Ascii character which can be used more easily in automatic analyses.

as *Intonation Functions*). I argued that this annotation is sufficient to account for all those aspects of prosodic representation which contribute directly or indirectly to syntactic interpretation and which can consequently be used to distinguish minimal pairs where intonation disambiguates syntactic ambiguity. The notion of syntactic interpretation is of course highly theory dependent (see discussion in Hirst 1977) but the resulting annotation scheme does provide at least a first approximation of a prosodic annotation scheme that does not directly refer to prosodic form.

INSERT FIGURE 1 ABOUT HERE

The original formulation of the IF annotation system was presented as a set of distinctive features. Most phonological frameworks today assume that phonemes are grouped into higher-level phonological constituents. Specifically, I assume here a hierarchy consisting of syllables, rhythm units, tonal units[4] and intonation units. With a hierarchical structure of this type only [± emphatic] and [± terminal] need to be retained specifically as features. The utterance above could then be transcribed as in Figure 2, which is formally equivalent to the linear IF transcription. Although this tree representation is no doubt closer to an appropriate linguistic representation of the prosody, the interlinear IF system is formally equivalent to the tree, and so I shall continue to use this more compact representation system in the rest of this paper.

INSERT FIGURE 2 ABOUT HERE

It is obvious that a complete functional representation of prosody would be much richer than IF annotation, which completely leaves aside whole areas of

[4] Both the **rhythm unit** and the **tonal unit** were originally proposed for the description of English prosody by Jassem 1952. For arguments in favour of the tonal unit as a phonological constituent for the intonation of English and French cf Hirst 1988, for the rhythm unit cf. Bouzon & Hirst 2004, Bouzon 2004.

prosodic expression including discourse structure, speech act, expression of affect etc.. It seems, however, a legitimate strategy to assume that the expression of affect, for example, will enrich, rather than override, the more linguistic role of prosody encoded with IF annotation, which can consequently be taken as a useful first approximation.

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 proposed (Hirst et al 2000) that the representation of prosodic form should involve a number of different levels including a level of phonetic representation, consisting of quantitative 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 more directly linked to the representation of prosodic function (see Hirst 1998 for application of this to the intonation of British English). In the next sections I outline briefly some specific proposals concerning these different levels of representation. Most of these specifically concern the representation of pitch and intonation since this is the area in which I have worked most systematically. I assume, however, that the same techniques can, and I hope will, be applicable to the representation of duration and intensity as well as to the comparatively little studied area of voice quality.

5.1 Phonetic representation

The MOMEL algorithm developped in Aix-en-Provence (Hirst & Espesser 1993, Hirst et al. 2000), provides an automatic phonetic representation of a raw fundamental frequency curve. The algorithm is sometimes referred to as a stylisation of fundamental frequency but it should more properly be called a model since it consists in factoring the 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), which we assume to be the essential component contributing to the linguistic function of the contour, and a microprosodic component consisting of deviations from the macroprosodic curve and caused by the nature of the current segment (voiced/unvoiced obstruent, sonorant, vowel etc) (cf Di Cristo & Hirst 1986). The output of the algorithm is a sequence of points <time, frequency>, referred to as target points, which are sufficient to define the macroprosodic component of the fundamental frequency when linked by a quadratic spline function (i.e. a continuous and smooth sequence of quadratic functions). These target points correspond approximately to the turning points of the continuous function; more exactly they correspond to the points where the first derivative (slope) of the function is zero. Note that since these target points are derived solely from the fundamental frequency curves, without access to any other linguistic information, the algorithm does not provide any indication as to the way in which the points are to be aligned with the prosodic structure of the utterance. The whole issue of the relation between the quantitative target points and the prosodic structure is an empirical one much in need of investigation.

Momel is currently available in a number of different implementations in various speech-analysis environments including Mes for Unix (Espesser 1996), SFS for Windows and Unix (Huckvale 2000-2005) as well as an external C function which can be called by script (Auran 2003) from within the multi-platform system Praat (Boersma & Weenink 2005)

A recent evaluation of the algorithm (Campione 2000) was carried out using recordings of the continuous passages of the Eurom1 corpus (Chan & al 1995) for five languages (English, German, Spanish, French, Italian), in all, a total of 5 hours of speech). The evaluation estimated a global efficiency coefficient (as calculated by the F-measure[5]) of 95.5% by comparison with manually corrected target point 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 less efficiency (93.4%) when applied to a corpus of spontaneous spoken French. The majority

[5] The *F*-measure is calculated as the harmonic mean (i.e. the product divided by the arithmetic mean) of the measure of *recall* (percent detected of total correct) and that of *precision* (percent correct of total detected). The *F*-measure is commonly used in the field of information retrieval as a global estimate of efficiency.

of these corrections involved systematic errors, in particular before pauses (especially preceded by a concave rising movement), which a recent improvement of the algorithm usually manages to eliminate.

INSERT TABLE 2 ABOUT HERE

5.2 Surface phonological representation

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. The reasonably theory-neutral (or at least "theory-friendly") nature of the modelling, however, 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) as well as the Fujisaki model (Mixdorff 1999).

The prosodic annotation alphabet INTSINT was originally 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 in this system are analysed as consisting of a sequence of tonal segments, defined in one of two ways. Three tonal segments are defined globally with respect to the speaker's current pitch range: **T**(op), **M**(id) and **B**(ottom). Three more are defined locally with respect to the preceding target point: **H**(igher), **S**(ame) and **L**(ower). We further assume an iterative variant of two locally defined tonal segments: **U**(pstepped), **D**(ownstepped), 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. An extension of the INTSINT system to annotate duration and timing has also been proposed (Hirst 1999) although this, as mentioned above, is an area which is still a subject of considerable speculation.

This transcription system, originally designed as a tool for linguists transcribing the intonation of utterances of different languages, was intended as a first approximation to a prosodic equivalent of the International Phonetic Alphabet. As we saw above this is specifically not the aim of 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). A simpler and more robust algorithm has since been developed (Hirst 2001). In this version, target points are coded on the basis of two speaker/utterance dependent parameters: *key* and *range*. Given these two parameters, the absolute tones are defined as the limits of the speaker's pitch range (**T** and **B**) assumed to be symmetrical around the central value (**M**) corresponding to the speaker's *key*. The relative tones are then defined by an interval between the preceding target point ($\mathbf{P_{i-1}}$) and the two extreme values (**T** and **B**) taken as an asymptote for these target points as in the following:

$$(5)\ P_i = P_{i-1} + c.(A-P_i)$$

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

This provides a very simple interpretation for the relative target points which are thus assumed to correspond to a tonal gesture moving either half of the way or a quarter of the way from the preceding target point towards the top or bottom of the speaker's current pitch range.

This algorithm was applied to the target points of the French and English passages of the Eurom1 corpus, and optimised over the complete parameter space (1000 iterations) defined by:

$$key = mean \pm 50 \text{ (in Hz by steps of 1 Hz)}$$

range *[0.5, 2.5]* (in octaves by steps of 0.1 octave)

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 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 3, 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.

During the optimisation process it is assumed that the speaker's *key* and *range* remain constant throughout the passage. This assumption, although fairly reasonable for the Eurom1 passages (as can be seen from figure 3), is obviously untenable for longer passages and in particular for spontaneous speech where changes in key and range are both frequent and communicatively significant.

INSERT FIGURE 3 ABOUT HERE

The problem is rather similar to that of the relationship between tempo and lengthening. Much interesting work remains to be done in this area, on the way in which listeners manage to differentiate long-term features like key, range and tempo from short-term features such as Top, Higher, Downstepped etc for tones or long/short for phonemes. Work in progress on the ProZed framework (Hirst 2001, Hirst & Auran forthcoming) is an attempt to provide a tool for analysis by synthesis, representing the two types of features as distinct characteristics which, we argue, is a necessary first step towards the automatic extraction of such representations.

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.

Example (4) above:

(4) | If you 'can't °lift it + 'ask !Peter to 'help you || (*IF annotation*)

would be converted to something like:

(5) If you can't lift it ask Peter to help you. (*INTSINT annotation*)
[M H B H M B T B B]

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

(6) If you can't lift it ask Peter to help you. (*target points*)
[135 163 95 143 135 95 191 95 95]

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

INSERT FIGURE 4 ABOUT HERE

Other 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 straightforward, although a number of somewhat arbitrary decisions have to be made as to what is the default interpretation of a given representation. Decisions also need to be made as to the manner in which the tonal segments are aligned with the prosodic structure, an area about which

comparatively little is as yet known. An implementation for a French text-to-speech system is described in Di Cristo et al (2000).

The inverse mapping, from INTSINT to IF, 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 a sequence of upstepping accents, with the first accent low and each subsequent accent slightly higher than the last. This corresponds to the global pattern which has sometimes been called "surprise/redundancy". This "upstepping head" has the effect of reinforcing the fact that the final fall is heard as 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 since it can occur with other nuclear patterns.

The fact that an upstepping head begins with a low accent may further be reinforced by a high onset for any preceding unstressed syllables (high pre-head). Once again, while this is a common characteristic of patterns with upstepping heads it is by no means necessary and it is again 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) but once again this is neither a necessary nor a sufficient characteristic of such patterns.

In (Hirst & Di Cristo 1984, Hirst 1998) I argue that surface phonological representations for non-emphatic and emphatic intonation patterns in French and 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:

(a) French, for example, like other Romance languages would be characterised as having a right-headed Tonal Unit (i.e. grouping unstressed syllables with a following stressed syllable) whereas English, like other Germanic languages would have a left-headed Tonal Unit (grouping unstressed syllables with a preceding stressed syllable).

(b) The underlying tonal template for French, in this analysis, is the sequence [L H] whereas in English the underlying sequence would be [H L]. These underlying templates can then be used to derived a typical set of surface intonation patterns for the two languages (Hirst & Di Cristo 1984).

(c) It seems furthermore that the organisation of the hierarchy of prosodic constituents might be different in English and French. In English the rhythm unit would seem to be best analysed as at a lower hierarchical level than the tonal unit, as originally suggested by Jassem (1952). In French, on the other hand, 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 would be no distinctiveness 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. Thus, French has nothing like English:

(7) a. |*The °sun's 'shining* ||
b. | *The °baby's 'crying* ||

According to this hypothesis, this type of de-accenting would only be possible in French in emphatic patterns and in patterns containing final postposed phrases. For these a different analysis would be necessary in English and in

French, making use perhaps of a recursive structure of embedded intonation units such as:

(8)a. *[[Nous partons tôt] demain matin]*
b. *[[We're leaving early] tomorrow morning]*

where the main pitch fall is likely to occur on the words *tôt* and *early* with *demain matin* and *tomorrow morning* being pronounced with low static pitch.

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 the 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 (2000) noted, 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, finally, suggests a strategy of analysis by synthesis which seems rather promising as a technique for providing an empirical justification for a more extended functional representation system.

In this approach, a preliminary IF annotation would be 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 far too 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.

Our group in Aix en Provence has recently completed an enhanced version of the Marsec corpus consisting of five and a half hours of continuous (mainly scripted) speech which is entirely transcribed both orthographically and phonetically and aligned with the speech signal (Auran et al. 2004). The original corpus was also entirely transcribed prosodically using the British school *tonetic stress mark* (TSM) system. Preliminary analysis suggests that the TSM system contains sufficient information to derive the equivalent of our IF annotation and this will consequently be an excellent test-bed to apply the strategy of analysis by synthesis described above. Since we are making the Aix-Marsec database freely available to the scientific community, this could also be an excellent opportunity to compare different annotation systems using the same data.

Acknowledgements

An earlier version of this paper was published in *the Festchrift for Professor Wu Zongji's 95th Birthday* (Foreign Language Teaching and Research Press, Peking 2003.) I have presented various parts at conferences and meetings in Joensuu, Finland; Belo Horizonte, Brazil; Tokyo, Japan; Nantes, France; Rome, Italy; Tianjin, China and Beijing, China and of course in Nara, Japan for Speech Prosody 2004. I should like to thank the participants and organisers of these meetings for giving me the opportunity of stimulating and fruitful discussions. My thanks also to two anonymous reviewers for their constructive comments and suggestions. It would be nice if I could blame these people for all the remaining inaccuracies and inconsistencies in my text but, alas, for these I have no-one to blame but myself.

Figures

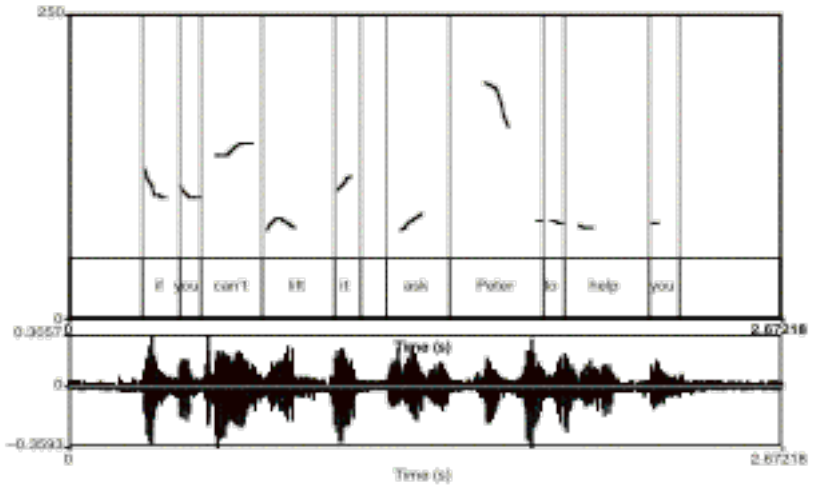

Figure 1. Signal and F0 of "If you can't lift it, ask *Peter* to help you."

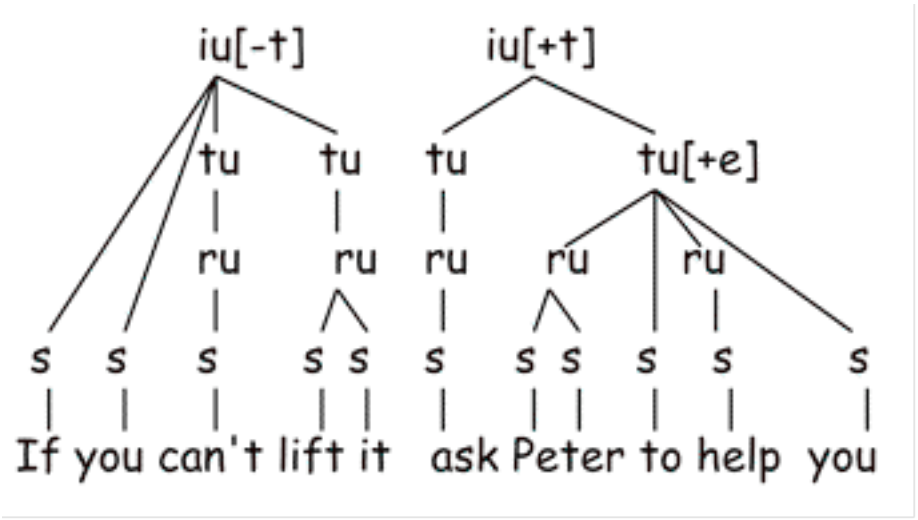

Figure 2: Representation of the functional prosody of example 4 as a hierarchical phonological structure with phonemes grouped into syllables (**s**), rhythm units (**ru**) tonal units (**tu**) and intonation units (**iu**).

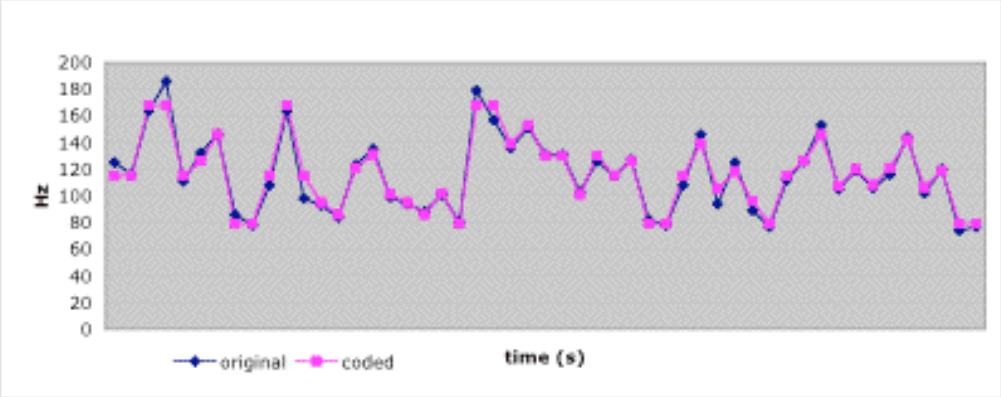

Figure 3. Coding of the F0 target points from a sample passage from the Eurom1 corpus showing the original target points estimated by the Momel algorithm and the target points derived from the optimised INTSINT coding.

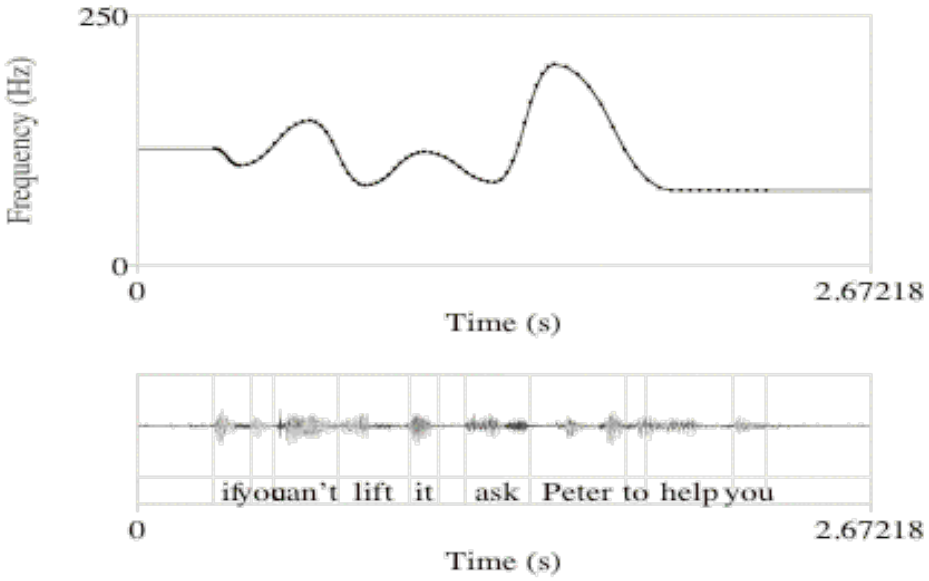

Figure 4. Continuous quadratic spline curve produced as output from the sequence of target pitches in example (6)

Transcriber agreement	
Presence vs. absence of edge tone	85-92%
Presence vs. absence of prominence	81-91%
Type of edge tone	for 6 of 9 labels < 50%
Type of pitch accent	for 6 of 8 labels < 50%

Table 1: Agreement between ToBI transcribers broken down into different categories. Data from Wightman 2002.

corpus	*language*	*Number of target points*			*Evaluation*		
		detected	added	removed	recall	precision	F
Eurom	*English*	8380	623	125	98.5	93.0	95.7
	French	6547	423	130	98.0	93.8	95.9
	German	13595	1145	506	96.3	92.0	94.1
	Italian	9475	337	330	96.5	96.4	96.5
	Spanish	8985	651	16	99.8	93.2	96.4
	All	46982	3179	1107	97.6	93.5	95.5
Fref	*French*	9835	532	744	92.4	94.5	93.4

Table 2. Statistical analysis of the efficiency of the Momel algorithm applied to the Eurom1 corpus (5 languages) as well as for a corpus of spontaneous French (Fref) . See footnote 4 for details on the statistics.

References

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

Auran, C; Bouzon, C. & Hirst, D.J. 2004. The Aix-MARSEC database. Towards an evolutive database for spoken British English. in *Proceedings of the Second International Conference on Speech Prosody*. Nara, March 2004, 561-564.

Boersma, P., & Weenink, D. 1995-2005. 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 *Proceedings of the Second International Conference on Speech Prosody*. Nara, March 2004, 223-226.

Bouzon, C. (2004) *Rythme et structuration prosodique en anglais britannique contemporain*. Doctoral thesis, Université de Provence.

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-to-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.

Grabe, E., Post, B. and Nolan, F. (2001). Modelling intonational variation in English: the IViE system. *Proceedings of Prosody 2000*, Adam Mickiewicz University, Poznan, Poland, 51–58.

Gussenhoven, C. (2002). Phonology of intonation. *GLOT International* 6 (Nos 9/10). 271-284

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. 2001. Automatic analysis of prosody for multi-lingual speech corpora. in E. Keller, G. Bailly, A. Monaghan, J. Terken & M.Huckvale (eds.) *Improvements in Speech Synthesis*. (London, John Wiley). 320-327.

Hirst, D.J. & Auran, C. forthcoming. Analysis by synthesis of speech prosody: the ProZed project.

Hirst, D.J. & Di Cristo, A. 1998. A survey of intonation systems. in Hirst & Di Cristo (eds) 1998, 1-44.

Hirst, D.J. & Di Cristo, A., 1984. French intonation: a parametric approach. *Die Neueren Sprachen*, 83 (5), 554-569.

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. 2000-2005. Speech Filing System. Tools for speech research.

http://www.phon.ucl.ac.uk/resource/sfs/

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

Jun, S.-A. 2005. *Prosodic Typology. The Phonology of Intonation and Phrasing*. Oxford, Oxford University Press.

Ladd, D. R. . 1996. *Intonational Phonology. (= Cambridge Studies in Linguistics 79)*. Cambridge: Cambridge University Press.

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.

Pierrehumbert, J. (2000) Tonal elements and their alignment, in M. Horne (ed), 11-26.

Scarnà A. and Ellis A. W. (2002) On the assessment of grammatical gender knowledge in aphasia: The danger of relying on explicit, metalinguistic tasks *Language and Cognitive Processes*, 17 (2), 185-201

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 *Proceedings of the First International Conference on Speech Prosody*, Aix en Provence April 2002.

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.