Levels of representation and levels of analysis for the description of intonation systems

Daniel J. Hirst, Albert Di Cristo, Robert Espesser

Merle Horne (ed) Prosody : Theory and Experiment. Kluwer Academic Press.

Levels of representation and levels of analysis for the description of intonation systems.

Daniel Hirst, Albert Di Cristo & Robert Espesser

Abstract

It is argued that a satisfactory global theory of intonation will require four levels of analysis : (i) physical (acoustic, physiological) (ii) phonetic (iii) surface phonological and (iv) deep phonological. The theoretical and cognitive status of each level is discussed and specific proposals are made for a model respecting such an overall architecture as well as a condition of interpretability which requires that each level of representation be interpretable in terms of adjacent levels. The level of phonetic representation is conceived of as providing an interface between abstract cognitive representations and their physical manifestations. This level is also assumed to provide an interface between constraints on production and perception. For fundamental frequency an algorithm, MOMEL, for the automatic derivation of a representation as a sequence of target-points is presented. The level of surface phonological representation is seen as the prosodic equivalent of the International Phonetic Alphabet for phonemic representation. A symbolic coding system for fundamental frequency patterns (INTSINT) is described which is currently being used for the automatic coding of fundamental frequency patterns for continuous texts in a number of different languages. The level of deep phonological representation is described as the level of linguistically significant choices which interact with a number of language-specific prosodic parameters to generate observed intonation patterns.

1. Introduction

The linguistic description of the intonation systems of different languages, like that of any other aspect of language, can be thought of as a rather indirect process of extracting linguistic information from measurable physical data. As has long been known, there is no automatic technique for performing this operation. As Chomsky (1964) has pointed out, there is no general 'discovery procedure' we can appeal to.

The development of large corpus-based studies and the introduction of widely available automatic modelling techniques do, however, bring the hope that our knowledge of these systems may increase significantly over the next few years. One of the reasons for this is that as we increase our database we are able to formulate and test more and more empirical predictions about the data. The comparison of the predictions with the observed data in turn leads us to formulate more constrained hypotheses about the nature of the phonological representations we hope to bring to light.

At the same time, the development of studies dealing with inter-speaker and intra-speaker variability will teach us more and more about the way in which prosodic systems can vary within the same language or dialect. Similarly it can be expected that the availability of comparable data derived from descriptions of a number of different languages or dialects will enable us to separate out with more and more confidence the language-specific from the universal characteristics of prosodic systems and this will allow us once more to formulate considerable constraints on the nature of these systems.

It must be emphasised that this continuous dialogue between empirical data and linguistic theory is at each step a process of formulating hypotheses on the basis of available data and of testing these on new data. All hypotheses involved are, naturally, only provisional and are liable to be questioned at all times. The higher the degree of abstraction of a hypothesis, however, the greater the quantity of data necessary before we call it into question and look for a better hypothesis to replace it. This means that the fact that different teams of researchers work with different theoretical backgrounds, far from being a handicap, is in fact a guarantee that research is not confined to what may after all turn out to be a blind alley.

In this paper we formulate a general picture of an overall phonological and phonetic description of intonation and we then make a number of specific proposals for implementing the different levels of representation. Much of this material has been presented elsewhere[1] but we are grateful to the editor for this opportunity to present an overall synthesis here.

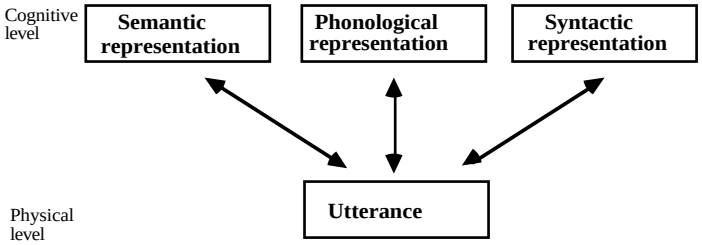
2. Levels of representation and levels of analysis

It is obvious that we need to distinguish at least two levels of representation (cf 't Hart & Collier 1975). At the most abstract linguistic level we want to be able to represent somehow the knowledge that a speaker needs to acquire when he learns a language. At the other extreme, we want to relate such a representation to the physical manifestations of this linguistic knowledge : the corresponding acoustic and physiological characteristics of utterances. Although the distinction between abstract linguistic representations and concrete physical representations is fairly uncontroversial, different approaches tend to differ in the relative importance which they attach to each of these two levels (cf discussion in Ladd & Cutler 1983).

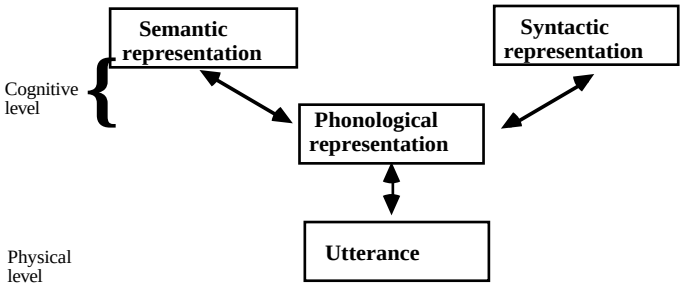
We can note that in these two cases the term 'representation' is not being used in strictly the same way. When we refer to linguistic representations, we assume that we are describing, however

imperfectly, the way in which speakers of the language themselves represent the information in their mind. With physical representations, we are rather describing the way in which we, as scientists, choose to analyse the data. In order to distinguish these two types of representation we can use the specific terms *cognitive representation* and *analytical representation*. In many cases of representations which are intermediate between the purely physical and the purely cognitive extremes, it is an empirical question whether they should be considered cognitive or analytical. Henceforth, when we refer simply to 'representations' without any qualifying term, it can be assumed that we are being deliberately vague as to which of the two meanings is intended. It could be argued that one of the principal aims of linguistic investigation is to provide a satisfactory theory of how analytical and cognitive representations are related.

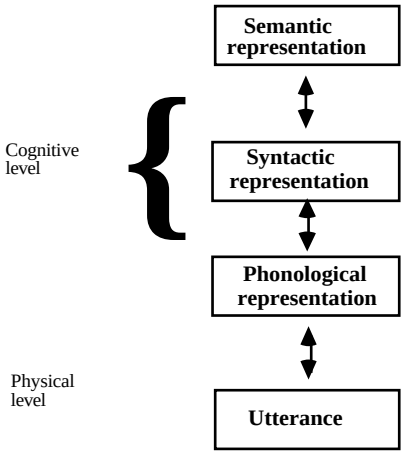
While the existence of an abstract cognitive representation is fairly widely accepted, there is little or no consensus as to its nature. For many linguists, linguistic representations are composite objects and the pronunciation of an utterance makes use of information from a number of different linguistic components : syntax and semantics as well as phonology as in the following diagram :



For others (cf in particular Selkirk 1978, 1984, 1986) there exists an autonomous level of phonological representation which encodes all (and only) the information concerning the pronunciation of an utterance. Schematically :



This hypothesis places a very strong constraint on the nature of linguistic representations and in the absence of convincing evidence against it we shall consequently assume such an organisation. In fact many linguists would place even stronger constraints and assume an organisation such as the following :



where the syntactic component (including the lexicon) is assumed to communicate with both the phonological and the semantic component but where there is no direct link between semantics and phonology. Nothing crucial in the rest of this chapter depends on the difference between this framework (that of the so-called 'Extended Standard Theory' of generative grammar Chomsky 1981) and the one mentioned previously. A lot depends on the precise role which the theory attributes to the semantic

component and its relation to pragmatic interpretation. There are a number of very important issues at stake here but which would obvously take us far beyond the scope of this chapter. For discussion see Hirst (1987, 1993).

It can be seen from the above diagrams that a phonological representation must fulfill two purposes : it must provide both the information necessary for the pronunciation of an utterance and the information necessary for its syntactic and semantic interpretation. This in fact provides us with a useful constraint which we shall refer to as the 'Interpretability Condition' and which states :

Interpretability Condition :
Representations at all intermediate levels must be interpretable at both adjacent levels : the more abstract and the more concrete.

There is nothing in what we have said so far which is specific to intonation. In the case of lexical items, for example, this framework embodies the insights of de Saussure 1916concerning the double nature of the linguistic sign as an arbitrary association between *signifiant* and *signifié*. In the case of prosody and intonation, we can usefully make a similar distinction between *functional representations* which encode the information necessary for the syntactic and semantic interpretation of the prosody of an utterance and *formal representations* which encode the prosodic information necessary for its pronunciation. Most systems of transcription for intonation mix functional and formal characteristics and it is of course an empirical question whether these should be encoded in separate representations or not. The fact that different languages make use of different prosodic forms for encoding the same prosodic functions seems to us evidence, though, that they should be separated : the inventories of prosodic forms and prosodic functions might then both be part of universal linguistic theory while the specific mapping between forms and functions in any given language would be defined by language specific parameters.

An example of such a difference is the distinction between lexical and non-lexical use of prosody in different languages. It has often been noted that there are no specific acoustic characteristics which distinguish languages which make use of lexically distinctive stress and tone from other languages which do not. Thus, there is no obvious acoustic cue which would allow us to distinguish stressed syllables in languages with 'free' or 'distinctive' stress (English, Russian etc.) from those in languages with 'fixed' or 'non-distinctive' stress (French, Polish, Finnish etc). Presumably the (surface) phonological representation of stress is the same in both types of languages.

Similarly, in tone languages (Chinese, Yoruba etc), we can assume that at least some of the melodic characteristics of utterances are lexically determined. It has however become standard in recent years to formulate phonological models of intonation which derive intonation patterns in non-tone languages from phonological tones, together with appropriate rules specifying how they are aligned with the accented syllables. This suggests the fascinating possibility that phonological representations in all languages draw from a universal set of prosodic characteristics which are either lexically specified or which are introduced in conformity with language specific parameters and which are subsequently mapped onto phonetic characteristics.

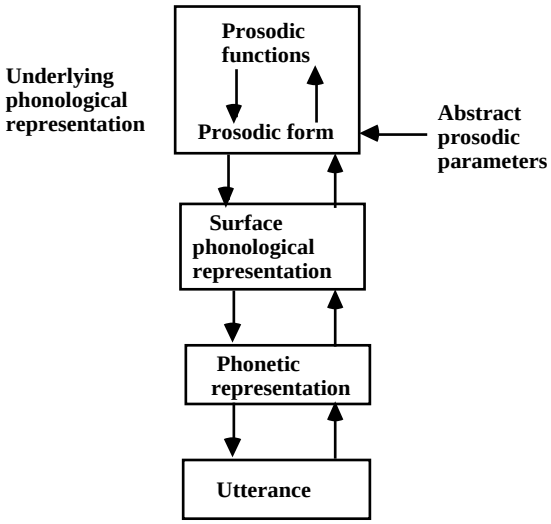
Once again it is an empirical question whether phonetic representation is a distinct level or whether, as suggested by Pierrehumbert & Beckman (1988), we should think of phonetics rather as a *dynamic interpretative process* mapping phonology directly onto acoustics and physiology. Whatever the theoretical status of phonetic representation, we believe that it is a useful heuristic strategy to postulate a distinctive level of phonetic representation which can then be mapped (following the interpretability condition) onto both physics and phonology. As Pierrehumbert & Beckman themselves put it : “the division of labor between the phonology and the phonetics is an empirical question, which can only be decided by constructing complete models in which the role of both in describing the sound structure is made explicit." (*op. cit.* p 5).

We can note that the term 'phonetic representation' has at times been used to cover a number of different phenomena which really need to be distinguished. Phonetics is sometimes used as a synonym for acoustics and physiology. It should be clear that we wish to distinguish these levels of analysis. Phonetics is also sometimes used as a synonym for 'surface phonology' as in the terms 'phonetic transcription' or the 'International Phonetic Alphabet'. For Trubetskoy (1949), the distinction between phonology and phonetics is one between discrete and continuous phenomena. In this sense then a 'phonetic transcription' would more appropriately be termed a 'surface phonological transcription'.

Between the underlying phonological representation and the physical representation we wish then to postulate two distinct levels : the level of surface phonology and the level of phonetics. The level of surface phonology is a level of distinctive discrete categories with which we can describe surface phenomena cross-linguistically. The level of phonetics is the level of continuously variable phenomena from which we have factored out universal constraints on the production and perception of sounds. We can illustrate these distinctions with the example of durational characteristics. Duration is often referred to as one of the three prosodic acoustic parameters, the other two being fundamental frequency and intensity. Unlike the other parameters, however, duration is not purely acoustic : it is impossible for a machine to produce something like a 'duration curve' in the same way that machines can produce intensity curves or

fundamental frequency curves. In order to measure duration we need to posit boundaries which are associated with phonological categories such as phonemes or syllables. Since duration is a continuously variable relationship between phonological units and physical parameters it fits precisely the definition we have given of a phonetic characteristic. Studies of the duration of phonological units of a large number of languages might well lead us to the conclusion that only some small finite number of durational distinctions are ever distinctive on a cross-language basis. This could then lead us to set up discrete surface phonological categories and we should then hope to be able to predict the observed range of values from an even more restricted number of underlying representations.

The description of the prosody of languages can thus be seen as a continuous process of defining representations at different levels together with constraints on these representations. The more we learn about representations at any one of the different levels we have described above, the more we shall know about other levels since we project supplementary constraints via the Interpretability Condition onto adjacent levels. The picture we have built up so far can be illustrated as follows :



Prosodic functions are mapped onto underlying prosodic forms in conformity with language specific abstract prosodic parameters. These underlying forms are mapped onto surface forms then onto phonetic representations before being output as physical correlates of utterances.

To summarise, we argue, then, that a satisfactory global theory of intonation will require four levels of analysis : (i) physical (acoustic, physiological) (ii) phonetic (iii) surface phonological and (iv) underlying phonological, the latter comprising both a representation of prosodic functions and prosodic forms.

In the rest of this chapter we present specific proposals for a model respecting the overall architecture we have just sketched. A number of the features of the model are yet to be developped in particular for the representation of durational and intensity characteristics as well as the exact nature of the alignment between the tonal categories and the segmental categories. This should in no way be taken as reflecting a lack of interest in these aspects but rather our current ignorance of the way in which these characteristics can most appropriately be integrated into represention at different levels.

3. Specific proposals.

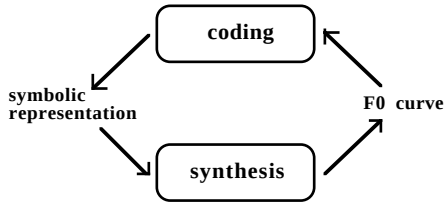
3.1. Phonetic representation.

The hypothesis behind the distinction between physical and phonetic levels of representation is that all languages obey the same physical constraints in the production and perception of variations in fundamental frequency, intensity and duration. A phonetic representation, then, is one where such universal constraints have been factored out (cf Di Cristo 1985 p 31).

Phonetic models of fundamental frequency patterns have been developped which attempt to account for physiological aspects of the production of intonation contours, as in the work of Fujisaki and colleagues (Fujisaki & Nagashima 1969; Fujisaki 1988, 1997). Other researchers, in particular those of the 'Dutch school' : 't Hart et al. 1990, Terken 1991, Terken & Hermes this volume as well as Rossi et al 1980, House 1990, d'Allessandro and Mertens 1995 etc, have concentrated on modelling more particularly the perceptual aspects of intonation patterns. If our conception of phonetics (as the level

where constraints on perception and production are resolved) is correct, then neither of these approaches is giving the complete picture : both aspects need to be incorporated into a more general model.

A number of different implementations of phonological/phonetic models of intonation have been designed to derive an acoustic output (F_0 curve) from a symbolic input (for an overview cf Hirst 1992). As in all fields of speech analysis, however, it is the inverse problem which is really the most challenging. Given an F_0 curve, how can we recover a symbolic representation? Even if we are able to perform such symbolic coding automatically, how should we validate the output of such a programme? One way would be to require the symbolic representation to be in such a form that it can be used as input to a synthesis system, the acoustic output of which can then be directly compared to the original F0 curve. The coding problem is thus directly related to the synthesis problem and in the rest of this section we shall reserve the term 'model' for attempts to solve both the coding problem and the synthesis problem together :



It is obvious that an automatic modelling system would be highly desirable for a number of reasons. An efficient algorithm would be extremely useful for collecting data for improving both speech synthesis and automatic speech recognition. Such a tool would also of course be extremely valuable for obtaining empirical evidence for testing phonological models of intonation and examining the variability in prosodic parameters across languages, dialects and individuals.

Despite its obvious interest, until quite recently the modelling problem had received less attention from workers in the field than might have been expected. An early treatment is Scheffers (1988) who described a technique for obtaining an automatic piece-wise linear approximation of an F0 curve. More recent approaches approaches are Geoffroi (1993), Jensen et al (1993), ten Bosch (1993), Bagshaw (1994), Taylor (1994), Morlec et al. (1995), as well as a numer of contributions to Sagisaka et al. (eds) 1997.

In the rest of this section, after discussion of some of the background and assumptions behind our work, we present an algorithm which has been developed in our laboratory in Aix-en-Provence.

Common to many proposals for fundamental frequency is the idea that microsegmental characteristics of curves should be factored out (Di Cristo & Hirst 1986) so that the resulting curve is similar to that found on a sequence of entirely sonorant segments as found in sentences likes "Molly may marry Larry." An F0 curve, then, is modelled as the superposition of two components : a micro-prosodic component caused by the characteristics of the individual phonematic segments of the utterance and a macroprosodic component reflecting the choice of intonation pattern for the utterance.

3.1.1 The microprosodic component

Di Cristo & Hirst (1986) describe an experiment in which nonsense syllables "bababa" and "vavava" were pronounced by a single speaker in three different contexts :

(i) “_____, c'est un mot.” (__, it's a word.)
(ii) “C'est un mot, _____.” (It's a word, ___)
(ii) “C'est un mot, _____?” (Is it a word, ____ ?)

The contexts were chosen to ensure that the nonsense words were pronounced with rising, low flat and high flat pitch patterns respectively. The fundamental frequency at the centre of the consonants of the nonsense words was compared to that of the mean of the surrounding vowels. Nicolas (1989) carried out the same experiment with four subjects (two male, two female), six voiced consonants (b d g v z j) two contexts (high, low) and four repetitions. The relationship between the F0 on the consonant and on the surrounding vowels was found to be fairly linear. Although a logarithmic regression in fact gave slightly better predictions than a linear one for most speakers and most consonants, the differences between linear and logarithmic predictions were very small. We conclude then that the raw f0 curve can be factored out into a macroprosodic component onto which is superimposed a (linear or logarithmic) microprosodic component.

3.1.2. The macroprosodic component

It follows from the above that the macroprosodic component of an F0 curve will be practically identical to the raw F0 curve observed for utterances consisting entirely of vowels and sonorant consonants, since these are known to have the smallest micromelodic effect. Observation of such curves shows that they obey two constraints: they are continuous and they are smooth. More technically, both the curve and its first derivative are everywhere continuous. The simplest function which obeys both these constraints is a quadratic spline function which we have used successfully for many years now for very close modelling of observed fundamental frequency curves and we have argued elsewhere (Hirst 1980, 1983, 1987, 1992, Hirst & Espesser 1993) that a function of this type allows us to treat a sequence of target points as an appropriate *phonetic* representation for F0 curves[2].

A spline function of degree *n* corresponds to a continuous sequence of polynomials of degree *n*, the derivatives of which, up to and including degree *n-1*, are everywhere continuous. Cubic splines are commonly used for interpolating values in a sequence of which only certain values are known. Quadratic splines, by contrast, interpolate monotonically between points of the curve at which which the first derivative is zero. They have the advantage that they are completely locally determined, since any given stretch of the curve between two target points is entirely determined by those points. This means that any erroneous targets will only affect the immediately adjacent sections of the curve. Quadratic splines are defined by a sequence of triples <*t,h,k*> where *t* and *h* define the time and frequency of the target point and where *k* defines the spline-"knot", that is the inflection point of the s-shaped transition between two target points. In the rest of this paper we assume for simplicity that these transitions are symmetrical[3], that is that the inflection point is always situated halfway between two adjacent targets[4].

Hans 't Hart (1991) has suggested that our attempt to synthesise F0 curves with parabolas rather than with straight lines is misguided since subjects cannot hear the difference anyway. This calls for a few comments.

Trivially, it is of course possible to approximate any complex function to an arbitrary precision by a sequence of straight lines. At the limit, one line segment per pair of F0 values will be exactly equivalent to the output of a quadratic spline function. This is not of course what 't Hart has in mind. Note however that the straight line interpolation which he compares to parabolas is not simply linear interpolation between target-points. Instead the interpolation is between horizontal plateaux so that a simple rising pattern which we would code with two target points :

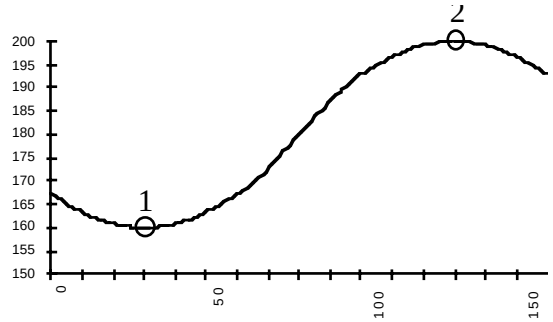


Figure 1 : Coding an F0 rise as a quadratic spline function with two target points.

would need to be coded as a sequence of five straight line sections as in :

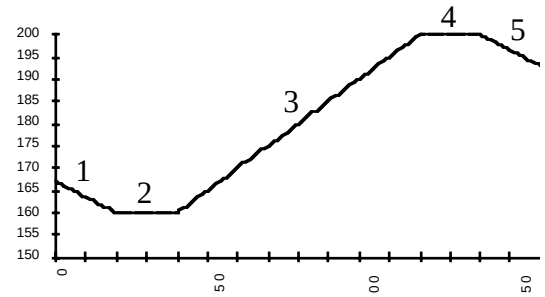


Figure 2 : Coding an F0 rise as a sequence of five straight lines.

The economy of such a representation is not evident.

We are not entirely convinced by 't Hart's argument that straight line synthesis will always be just as good as that using quadratic splines. Since quadratic splines give a closer approximation to real F0

curves than do straight lines, it seems possible that under certain circumstances these differences will be noticeable. Even the fact that subjects are apparently unable to distinguish certain stimuli does not prove that their reactions to these stimuli will always be identical under all conditions.

We might conclude that the very fact that an automatic coding algorithm such as that which we describe below works at all is a good enough reason for using the same model for synthesis since, as we suggested above, this means that the output of the analysis feeds directly into the synthesis. It would of course be possible to adapt the output to feed into staight line synthesis system, giving an output like that of Figure 2 for example[5], but then of course the direct comparison between observed F0 and modelled F0 would no longer be possible.

Finally, we entirely agree with 't Hart's finding that listeners are much more sensitive to variations in the temporal and frequential values of target points than they are to the nature of the transitions between targets. We take this, however, far from being a weakness of our system, as a strong argument in favour of using a sequence of target points as a phonetic representation of the F0 curve.

3.1.3. MOMEL : an algorithm for automatic modelling of F0 curves

The MOMEL algorithm which we present here consists of four stages. The central part of the algorithm (stage 2) is an asymmetric version of what we call *modal regression* as defined below. This stage works on the assumption that the only effect of the microprosodic component of a fundamental frequency curve is to *lower* the values of the underlying smooth continuous macroprosodic curve. The only case where microprosodic effects actually raise the F0 curve are at the boundary between voiced and unvoiced segments. We consequently implement a preliminary stage (stage 1) to make sure that any potential raising effects have been eliminated. The second stage then applies the modal regression technique within a moving window, to provide one optimal estimate of a local fundamental frequency target centred on each value of the fundamental frequency curve. The next stage of the algorithm (stage 3) selects a partition of these target candidates. The final stage (stage 4) reduces the candidates within each segment of the partition to a single target.

The technique used in this algorithm is one which we call 'modal regression' since it bears in fact the same relation to ordinary regression as the estimation of the mode of a distribution bears to that of the mean. Both the *mean* and the *mode* of a distribution are in some sense the values which are the closest to all the items of the distribution. In the case of the *arithmetic mean* this 'closeness' is defined by calculating the square of the distance from each item of the distribution and selecting a value such that the sum of these squared distances is minimal. By contrast, the *mode* of a distribution can be defined as the value which is less than a given threshold Δ from the largest possible number of items of the distribution. While a distribution has only one mean, it may have a number of different modes, and even when there is only one mode, its value may vary in function of the value of the threshold Δ.

Ordinary regression is basically the same as the calculation of the mean except that instead of comparing the items of a distribution to a single value, the items of a series are compared to the values of a well-defined function (in the case of linear regression to a straight line, in the case of quadratic regression to a parabola etc), the parameters of which are selected to minimise the sum of the squared distances from the individual items. By contrast, we can define *modal regression*, as selecting the parameters of a given function such that the values of the function are less than a given distance Δ from the largest possible number of items of a series. Note that while this definition tells us what the *aim* of modal regression is, it does not provide us with a procedure for selecting the parameters to fulfill this aim. The task is further complicated by the fact that, as with the mode of a distribution, there may be more than one value for the optimal parameters of a modal regression. One method for estimating these parameters is presented below.

In the case of fundamental frequency curves, we have suggested that with the exception of some local effects linked to the onset and offset of voicing and which the first stage of the algorithm described below is designed to eliminate, all other microprosodic effects consist of a lowering of the 'underlying' macroprosodic curve which we model as a quadratic spline function. We consequently introduce the further asymmetric constraint that the quadratic spline function we wish to find is such that there are *no* values more than a distance Δ above the function and as few values as possible more than the same distance Δ below it.

The four stages:

(1) preprocessing of f0
(2) estimation of target-candidates,
(3) partition of candidates,
(4) reduction of candidates

are described in the rest of this section. The typical values given for each parameter of the algorithm were obtained by a process of optimisation which is described in 3.1.4 below.

(i) preprocessing of f0

All values more than a given ratio (typically 5%) higher than both their immediate neighbours are set to 0. Since unvoiced zones are coded as zero, this preprocessing has essentially the effect of eliminating one or two values (which are often dubious) at the onset of voicing, i.e. about 10 to 20ms.

(ii) estimation of target-candidates

The following steps are followed iteratively for each instant *x*

a. Within an analysis window of length *A* (typically 300ms) centred on *x*, values of F0, (including values for unvoiced zones) are neutralised if they are outside of a range defined by two thresholds *hzmin* and *hzmax* and are subsequently treated as missing values. The threshold *hzmin* is a constant set to 50 Hz and the adaptative threshold *hzmax* is set to the mean of the top 5% of the F0 values of the sequence multiplied by 1.3.

b. A quadratic regression is applied within the window to all non-neutralised values.

c. All values of F0 which are more than a distance Δ *below* the value of F0 estimated by the regression are neutralised. (typical value of Δ fixed at 5%).

Steps **b** and **c** are re-iterated until no new values are neutralised.

d. for each instant ***x*** a target point <*t,h*> is calculated from the regression equation :

$$\hat{y} = a + bx + cx^2$$

where $t = -b/(2c)$ and $h = a + bt + ct^2$

If *t* is outside the current analysis window (i.e. if it is less than $x-(A/2)$ or greater than $x+(A/2)$) or if *h* is less than *hzmin* or greater than *hzmax*, then *t* and *h* are treated as missing values.

Steps **b, c** and **d** are repeated for each instant *x*, resulting in one estimated target point <*t;h*> (or a missing value) for each original value of Fo.

(iii) partitionning of target candidates

The sequence of target candidates is partionned by means of another moving window of length R (typically 200 ms) in which the average value of the targets in the first half of the window is compared to the average value in the second half. The boundaries of the partition are then taken as those values which correspond to a local maximum for this distance and which is greater than the overall average value of the distances.

Specifically *dt(x)* and *dh(x)* are calculated as the absolute mean distances between the *t* and *h* values of the targets in the first half of the window and those in the second half of the window. A combined distance *d(x)* is then obtained by weighting these distances:

$$d(x) = \frac{dt(x) * wd + dh(x) * wh}{wd + wh}$$

where:

$$wd = \frac{1}{mean(dt(x))} \text{ and } wh = \frac{1}{mean(dh(x))}$$

The boundaries of the partition are consequently set to each value *x* respecting the following three conditions :

$d(x) > d(x-1)$

$d(x) > d(x+1)$

$d(x) > mean(d(x))$

(iv) reduction of candidates

Within each segment of the partition, outlying candidates (for which either *dt(x)* or *dh(x)* are greater than one standard deviation from the corresponding mean values for the segment) are eliminated. The mean value of the remaining targets in each segment is then calculated as the final estimate of *t* and *h* for that segment.

The following figures illustrate the application of the algorithm to a sentence taken from the corpus Eurom1-French.

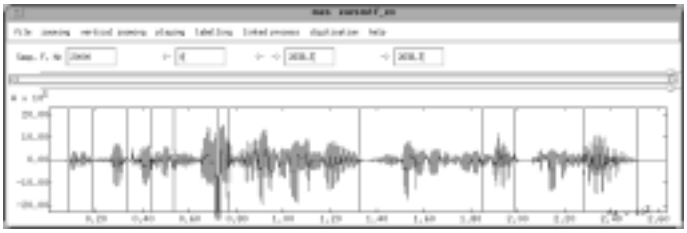


Figure 3 : acoustic wave form for the sentence "Il faut que je sois à Grenoble Samedi vers quinze heures." (I have to be in Grenoble by Saturday around 3 p.m.) taken from the French recordings of EUROM1. Vertical lines correspond to word labels time-aligned to the beginning of each word.

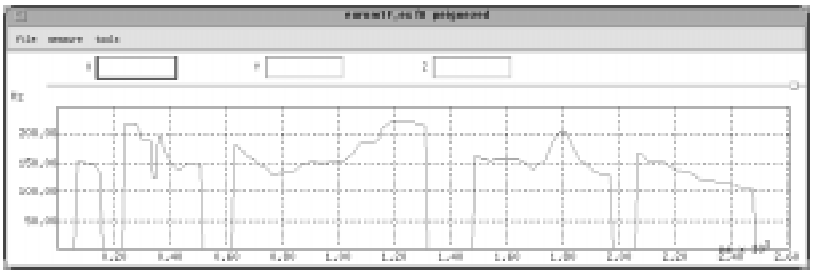


Figure 4 : fundamental frequency for the same sentence as in Figure 3, detected using a comb function.

In figure 5, the different target candidates estimated by the algorithm are represented by a grey line joining the centre of the analysis window (on the abscissae) to the x,y value of the target estimated for that window.

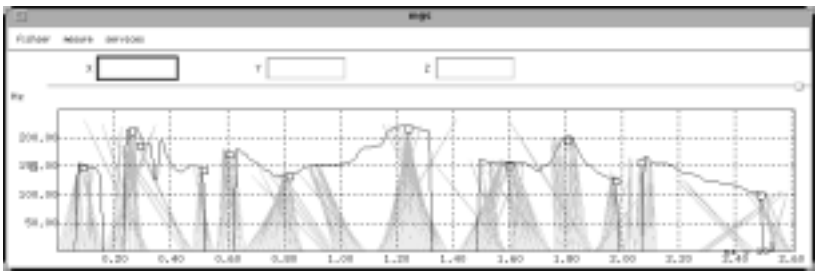


Figure 5 : Output of the automatic modelling program. Each target point candidate is visualised by a grey line connecting the centre of its analysis window on the x-axis to the target point. These candidates are then partitioned and the final target points selected are represented by the squares.

The squares in Figure 5 correspond to the final estimates of the targets after partionning the different candidates. The resulting modelled curve is illustrated in Figure 6 by the continuous smooth line.

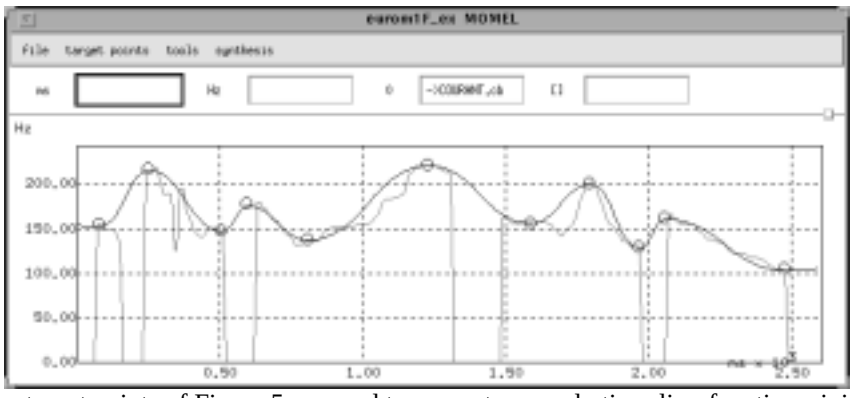


Figure 6 : The target points of Figure 5 are used to generate a quadratic spline function giving a close fit to the original curve.

While the algorithm described above is still somewhat less than perfect, it does seem to constitute at least a reasonable first approximation to a working phonetic model of fundamental frequency curves, incorporating both the coding and the synthesis of such curves. Since its development, the model has been used for the analysis of fundamental frequency curves in a number of different languages including English, French, Spanish, Italian and Arabic (Hirst et al. 1993) and is apparently fairly robust. Preliminary results obtained within the MULTEXT project (Véronis et al. 1994, Strangert and Aasa 1996, Campione et al. 1997) show that the modelling technique gives quite satisfactory results for English, French, Spanish, German and Swedish. Tests with a number of other Western and Eastern European languages are also in progress.

3.1.4. Evaluation of the phonetic representation

Fundamental frequency was detected every 10 ms using a combination of three methods : a comb-function (Martin 1981, Espesser 1982), AMDF and autocorrelation. No manual corrections were made to the detected F0 values.

The algorithm described above uses 3 independant parameters :

- analysis window *[A]*
- distance threshold *[D]*
- reduction window *[R]*

In order to optimise these parameters, a small corpus was used (*corpus VNV*) consisting of two sentences, containing all the stops and fricatives (and hence all the different microprosodic configurations) of French, spoken by ten subjects (5 male, 5 female).

S1 : La pipe de Jean s'est cassée en tombant de ta gabardine.
(John's pipe broke when it fell from your raincoat).

S2 : La fille de Charles Sablon a voulu un petit chien en guise de cadeau.
(Charles Sablon's daughter wanted a little dog as a present).

The following criteria were adopted :

(a) subjective evaluation

The original fundamental frequency curve and the modelled curve were compared visually. The number of manifest errors consisting of either missing targets or false targets was counted. The original recordings were compared (informally) with the same recordings resynthesised using the SOLA/PSOLA technique (Roucos & Wilgus 1985, Hamon, Moulines & Charpentier 1989) in order to check the relevance of the visual analysis.

(b) objective evaluation

A mean distance was calulated between the original fundamental frequency curve (hz_i) and the modelled curve (hz'_i) :

$$d = \frac{1}{n} \sum_{i=1}^{n} \left| 1 - \frac{hz'_i}{hz_i} \right|$$

During the optimisation of the algorithm a good correspondance between the two types of evaluation was observed.

The minimum value for F0 was found to be quite robust so that it was possible to fix the same value for all ten speakers *[hzmin]* = 50

The maximum value */hzmax/* was fixed as the mean of the top 5% of the speaker's F0 values multiplied by 1.3.

For the other three parameters the following values were found optimal for the corpus :

- *[A]* : 300 *[D]* : 5% *[R]* : 200

The algorithm was subsequently applied with its parameters fixed as above to four other corpora.

- *ATOME* : one continuous text in French read by 4 speakers (2 male, 2 female).

- *EUROM1*, English, French and German. For each language, 10 speakers (5 male and 5 female) read either 15 (English) 10 (French) or 20 (German) passages from a total of 40 different passages each consisting of 5 semantically related sentences (Chan *et al*. 1995)

The following table summarises the results of the analysis for the five corpora.

Table 1 : summary of distance measurements as defined above, number of target points detected and total duration for four different corpora.

corpus	mean distance	number of targets	total duration
VNV	6.22 %	284	49 s
ATOME	6.28 %	654	220 s
EUROM1 French	6.00 %	6747	2190 s
EUROM1 English	5.60 %	8680	2635 s
EUROM1 German	4.66 %	13995	4419 s

The mean distances recorded in Table 1 can be attributed to a number of different causes :

- erroneous detection of F0 (resulting in erroneous target points).
- errors of the algorithm itself i.e. targets either missed, added or displaced.
- microprosodic effects.

Microprosodic effects, according to preliminary estimates, probably account for no more than about one fifth of the distances recorded in Table 1 so that there is obviously still room for improvement of the algorithm.

Subjective evaluation of the application of the MOMEL algorithm to the EUROM1 passages for a number of different languages is currently in progress. Evaluators were asked to note cases where they could hear a difference between the original recording and the versions resynthesised from the target points, restricting attention as far as possible to differences in intonation. The results of this evaluation, although not amenable to statistical analysis are expected to provide us with an extremely valuable "error database" which can be used for the evaluation of future versions of the algorithm..

3.1.5 Remaining problems of phonetic representation.

As we mentioned above, we present in this section proposals for a phonetic representation of fundamental frequency curves. This representation is obviously far from constituting a complete phonetic model of intonation. Even in the domain of fundamental frequency the model as presented above does not address an important issue: that of the relative scaling of the target points - in particular when comparing data from different speakers. The problem of normalisation of acoustic data is one which needs to be addressed seriously if we hope to abstract away from speaker-specific variability. Prosodic characteristics differ from other phonetic properties in that what is important is generally not the absolute value of the property but its relative value. What is perceived as loud, long or high for one speaker may in fact be less loud long or high than what is perceived as soft, short or low for another speaker.

Recent research into the inter-speaker variability of prosodic characteristics (cf Campbell1992, this volume, Hermes & Van Gestel 1991, Terken & Hermes this volume) has provided a number of scaling techniques as a means to obtain speaker independent representations of prosody. These techniques provide a useful way of factoring out universal constraints on production and perception and should consequently provide a useful step in the process of extracting a phonetic representation from acoustic data.

3.2. Surface phonological representation.

The hypothesis behind the distinction between surface and deep phonological representations is that there is a level of description at which we can describe the prosody of different languages cross-linguistically using a restricted inventory of symbols much in the way that the International Phonetic Alphabet is used to describe the vowels and consonants of different languages.

Our examination of intonation patterns described for twenty different languages (Hirst & Di Cristo in press) led us to the conclusion that there is indeed a useful level of generalisation at which we can describe intonation patterns across languages on a surface level.

3.2.1. INTSINT : an INternational Transcription System for INTonation.

We proposed, as a first approximation, a transcription system (INTSINT) by means of which pitch patterns can be coded using a limited set of abstract tonal symbols, {T,M,B,H,S,L,U,D} (standing for : Top, Mid, Bottom, Higher, Same, Lower, Upstepped, Downstepped respectively). Each symbol characterises one of the target points of the phonetic representation. Unlike other systems of representation such as ToBI (Silverman et al., Pierrehumbert this volume), the use of INTSINT does not require that the inventory of pitch patterns of a given language alreadybe known before transcriptions can be made. It can thus be used, like the International Phonetic Alphabet, as a tool for gathering data on the basis of which phonological descriptions may be elaborated.

The rationale behind the INTSINT system is that the Fo values of pitch targets are programmed in one of two ways : either as **absolute** tones {T, M, B} which are assumed to refer to the speaker's overall pitch range (at least within the current Intonation Unit), or as **relative** tones{H, S, L, U, D} assumed to refer only to the value of the preceding target point. For relative tones, a distinction is made between **non-iterative** {H, S, L} and **iterative** {U, D} tones, since in a number of descriptions it appears that iterative raising or lowering uses a smaller F0 interval than non-iterative raising or lowering. The tone S has no iterative equivalent since it would be impossible to locate intermediate S tones temporally in a sequence of such tones. For illustrative purposes we propose iconic symbols which can be used to align the tonal symbols with the orthographic transcription. The following table gives the orthographic code followed by the recommended iconic symbols (all of which are taken from the widely available font *Symbol*):

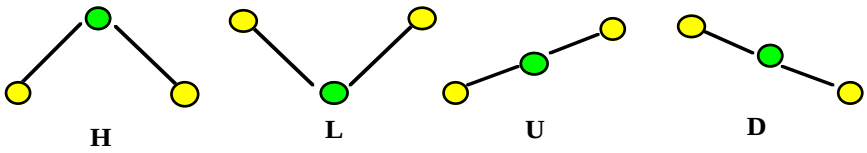
Table 2 : Orthographic and iconic symbols for the INTSINT coding system.

		Positive	***Neutral***	***Negative***
ABSOLUTE		T ⇑	M ⇒	B ⇓
RELATIVE	*Non-Iterative*	H ↑	S →	L ↓
	Iterative	U <	•	D >

With such a coding, an explicit statistical model can be made of the F0 patterning of a given corpus and it is possible to define empirical criteria for deciding whether a given transcription is better than another by comparing the mean of square deviations of the modelled data from the observed data. The adequacy of the symbolic coding can be further tested by regenerating target points (and hence Fo curves) from the coded representations. Using PSOLA resynthesis, the adequacy of the modelled utterance can be compared to that of the original recording. The fact that explicit statistical models can be compared empirically suggests of course the possibility that an optimal coding of a given speech segment could be derived automatically. We present a first approximation to such an algorithm in the following section :

3.2.2. An automatic coding algorithm for INTSINT.

The following algorithm represents a first approximation to such an optimal coding :
(i) the highest and lowest target values in the utterance are coded respectively T and B.
(ii) the first target point, as well as any which follow a silent pause of more than 250ms, is coded M (unless already coded T or B).
(iii) all other target points are coded with relative tones. A target which is less than a given threshold Δ Hz from the previous target is coded S. Otherwise it is coded H, L, U or D according to its configuration with respect to the preceding and following target points as in the following diagram :



Target points before a pause (>250 ms), where there is no relevant following target point, are (somewhat arbitrarily) coded as either S, H or L depending on the previous target.
(iv) the statistical value of each category of target points is then calculated : for absolute tones the mean value is taken, for relative tones a linear regression on the preceding target is calculated.
(v) Any target points originally coded H or L can be recoded as T, U, B or D if this improves the statistical model (i.e. the value predicted from the recoded tone is closer to the observed value than that predicted from the original coding).
(vi) steps (iv) and (v) are then repeated until no more points are recoded.

The fact that only H and L targets are allowed to be recoded in step (v) ensures that the algorithm converges to a (local) optimum.

The sample sentence analysed in section 3.1 above coded in this way corresponds to the following transcription :

(1) Il faut que je sois a Grenoble Samedi vers quinze heures
⇒ ⇑ ↓ ↑ ↓ ⇑ ⇒ ⇑ ↓ ↑ ⇓

There are, of course, a great number of other possible statistical models which need to be tested. The simple model we have described above seems however to give quite good results for the material to which it has been applied (i.e. isolated sentences and short passages for a number of different languages) and it is now a question of collecting more empirical evidence in order to see how such a model can best be improved.

3.2.3 Preliminary results for surface phonological representations.

The algorithm described in §3.2.2 was applied to the EUROM1 passages for both English and French. The output of the algorithm is two coefficients *a* and *b* for each tone and each reading of a passage which can then be used to model the value of a given target point P_i with the linear equation :

(2) $P_i = a.P_{i-1} + b$

For 'absolute tones' (**T**, **M** and **B**) the coefficient *a* was set to 0 so that the tones were modelled as a constant (the mean for the tone). For 'relative' tones (**H**, **S**, **L**, **U**, **D**) the coefficients were derived by linear regression on the preceding value. It was observed that in over 75% of the cases the coefficient *a* was estimated as a constant with a value between 0 and 1[6]. This suggests that a more constrained model might be appropriate since when a is between 0 and 1 the linear equation (2) applied re-iteratively converges to an asymptotic value[7] :

(3) $A = b/(1-a)$

so that equation (2) could be replaced by the following :

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

In other words each time a lowering or raising factor is applied, this is the equivalent of going a fixed ratio of the distance remaining between the previous target and the asymptotic value. The following table gives the estimated mean or asymptote for male and female speakers split by language :

Table 3 : mean (T,M,B) or asymptote (H, U, S, D, L) values calculated for male and female speakers split by language for the EUROM1 passages analysed with the MOMEL and INTSINT algorithms.

sex:	**M**		**F**	
language:	**English**	**French**	**English**	**French**
T	205	200	327	377
H	160	191	287	310
U	184	232	195	336
M	162	140	261	261
S	106	78	284	343
D	53	54	143	215
L	93	111	161	206
B	92	98	139	179

Analysis of variance showed, as might be expected, highly significant effects for both tone and sex ($p < 0.0001$). Significant interactions between tone and sex ($p=0.039$) and tone and language ($p=0.049$) were also observed but the number of speakers (5 per sex per language) was obviously far too small for us to propose any interpretation of these results. Ignoring these interactions the following figure illustrates the means or asymptotes of the different tones expressed in ERB units (cf Terken & Hermes this volume) offset to the mean value of all the tones for male speakers and female speakers respectively :

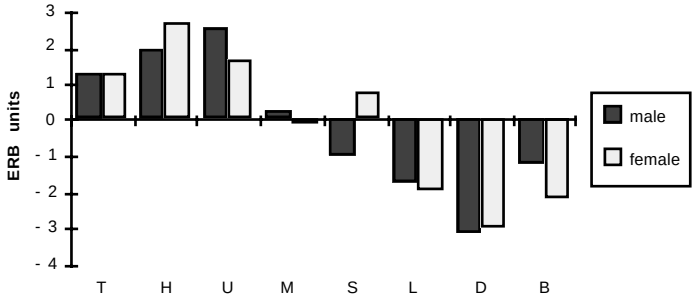


Figure 7 : Mean or asymptote values for each tone expressed in ERB units offset to the overall mean for male and female speakers respectively.

Hirst et al. (1991) suggested the possiblity that the same constants might be used for modelling both absolute tones and relative tones. Under this interpetation, T and H might share one target value while B and L might share another value[8]. In the case of absolute tones the target is actually reached whereas in the case of relative tones the target is simply approached. The data illustrated in Figure 7 seems compatible with this hypothesis. It is also compatible with the observation that the relationship between the two constants defining the upper and lower target values is close to that of one octave.

Making a number of simplifying assumptions, the whole system could in fact be reduced to a small number of speaker dependent variables :minimally a single parameter related to a speaker's physiological characteristics. It remains, however, to be seen how such a minimalist prosodic model can be related to the sort of inter and intra speaker variability discussed above (§1).

3.2.4. Remaining problems of surface phonological representation.

One obvious weakness of the model as presented above is the fact that it does not take into account the way in which 'absolute' tones T, M and B are themselves scaled throughout a more extended discourse. Nicolas (1995), Nicolas & Hirst (1995) showed that for continuous texts in French, appreciably better results were obtained if some scaling of this sort was introduced. In particular a T tone near the beginning of an intonation unit tended to be significantly higher than average in the first Intonation Unit of higher-level pargaraph (or "paratone") type structures and significantly lower in the last Intonation Unit of these structures. Final B tones were also significantly lower at the end of units of this type.

Since the scaling is assumed to apply across Intonation Units, one possible solution would be to use the same set of INTSINT symbols to code not only the individual target points, but also the overall range and register of the Intonation Units somewhat in the way recommended by the IPA working group on suprasegmental transcription (Bruce 1989). Assuming that Intonation Units are delimited by square brackets and that symbols within the square brackets define target points whereas symbols in round brackets define the pitch range of the following Intonation Unit a sequence such as example (1) above could be transcribed :

(5)a. Il faut que je sois a Grenoble Samedi vers quinze heures

(⇑)[⇑ ↓ ↑ ↓ ⇑](>)[⇑ ↓ ↑ ⇓]

where the intial (⇑) indicates that the pitch range of the first intonation unit is set to an initial high (Top) value, and the symbol (>) indicates that the entire pitch range of the second Intonation Unit is scaled somewhat lower than the first. Another possibility would be for the pitch range as described for (5a) to be the default values so that the sentence could be simply transcribed :

(5)b. Il faut que je sois a Grenoble Samedi vers quinze heures

[⇑ ↓ ↑ ↓ ⇑] [⇑ ↓ ↑ ⇓]

where by default the pitch range of the first Intonation Unit is set to Top and that of subsequent unmarked Intonation Units is Dowstepped.

Needless to say, we have not yet developped an algorithm which will convert a representation like (4) above into one like (5a) or (5b). Dividing a signal up into Intonation Units is a problem which is far from trivial and a lot of work remains to be done in this area. Introducing global parameters on Intonation Units such as in the above example makes it possible to produce much more accurate modelling of data but at the expense of introducing a number of extra degrees of freedom with the result that it is not a simple matter to develop heuristics to choose between possible alternative codings of a given surface realisation.

3.3. Underlying phonological representation.

The level of deep phonological representation can be thought of as the level of linguistically significant choices which interact with a number of language-specific abstract prosodic parameters to generate observed intonation patterns.

We made a distinction above between functional and formal representations of prosody and we mentioned that most transcription systems combine aspects of the two. Thus for example the 'tonetic stress' marking system of the 'British School' (cf. Cruttenden 1986) uses symbols such as {´; `; ^; ˇ} to indicate both the fact that a given word is highlighted, and the direction of the pitch-movement initiated on the stressed syllable of that word. Similarly the ToBI system (Silverman et al. 1992) makes use of symbols such as H* L% corresponding to a high tone associated with an accent followed by a low tone associated with a boundary. The symbols * and % encode functional aspects of the transcription whereas

the symbols H and L encode formal aspects. One purely functional system of representation of intonation is ordinary punctuation: readers are surprisingly consistent in the way in which they interpret punctuated texts even though there is absolutely no formal relationship between the punctuation symbols and their corresponding intonation patterns[9].

It has long been known that intonation has "a bewildering variety" of different functions (Ladd 1993 p 12.). Ladd mentions "signalling focus, emphasis, phrasing distinctions, lexical distinctions, speaker attitude, and many more" (op. cit.). If we restrict our attention to the more strictly linguistic functions of intonation, following Bruce (1985) we can distinguish two major functions : a weighting function (foregrounding/backgrounding) and a grouping function (coherence/boundary signalling). Work on prosodic structure theory has suggested that a fruitful way of thinking of prominent syllables is as the head of a prosodic constituent (Selkirk 1978, 1984, 1986). Halle & Vergnaud (1987) suggest that both prominence and grouping can be reduced to a single relationship of 'government' defined as the conjunction of two 'conjugate' structures : the pure representation of heads (prominence) and the pure representation of domains (grouping) - either can be recovered from the other provided that we specify by a language-specific abstract parameter the left-headed or right-headed nature of metrical constituents (p16)[10].

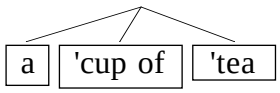
Representations of this sort entail the prediction that in addition to the distinction accented/unaccented, there can be different positions of the constituent boundary, depending on whether the head is taken to be at the beginning or the end of the prosodic constituent. An example of such a distinction can be made between accent groups in English and French. A number of facts are consistent with the idea that accent-groups in French are "right-headed" - that is they culminate in a accented syllable rather than starting with one as in English and presumably many other Germanic languages (Wenk & Wiolland 1982; Fant et al. 1991)[11].

Assuming that this is an appropriate distinction between English and French at least would mean that phrases like :

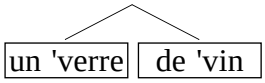
(6)a. a cup of tea
(6)b. un verre de vin (a glass of wine)

which presumably have the same syntactic structure in English and French, would be structured differently at the prosodic level as in the following:

(7)a .



(7)b.



Holmes (1996) in a contrastive study of English and French monologues observed a significant difference in the distribution of pauses in the two languages : English speakers may regularly insert pauses within a syntactic phrase whereas French speakers very rarely do so. When French speakers do pause after an article or a preoposition, they typically repeat the article or preposition when they resume speaking so that whereas an English speaker might produce say something like:

(8)a. a … cup of … tea

a French speaker is more likely to say:

(8)b. un… un verrre de… de vin.

In terms of prosodic structure these observations can be re-interpreted as saying that speakers tend more to pause between accent groups rather than within them.

The left/right headed nature of the stress-group is then an example of the sort of abstract prosodic parameter which we assume underlies the variability observed across different intonation systems. In the remainder of this section we present a sketch of the way in which we account for the differences between the basic intonation systems of English and French[12]. Since this sketch draws on characteristics of French which are not necessarily widely known we introduce this comparison by a brief account of French non-emphatic accentuation.

French is traditionally described as having systematic word-final accentuation (Halle & Vergnaud 1989). Recent studies have shown, however, that the actual accentuation of utterances in spoken French is considerably more complex than this (Hirst & Di Cristo 1984, Pasdeloup 1990, Di Cristo & Hirst 1997). In particular, pitch prominence is often given to the initial syllable of a word as well as, or instead of, to

the final syllable. This initial prominence, which in the past has often been taken for some type of emphatic stress, in fact occurs quite systematically in the speech of many speakers of modern French without conferring any particular emphatic connotation to the word. Other more conservative styles of French do not give prominence to the initial syllables of words.

There have been a number of attempts to account for the "probabilistic" nature of French accentuation (Fónagy 1980). A fairly satisfactory first appoximation can be obtained by the following rules which can be assumed to apply from left to right :

(i) divide the utterance into intonation units
(ii) assign an accent to the final syllable of the intonation unit.
(iii) assign an accent to (the initial and) final syllable of each accentable word

where the bracketed part of (iii) only applies in the less conservative styles mentioned above. Rule (iii) obviously considerably overgenerates accents. A further principle, similar to the well known "stress clash" rule in English and many other languages limits the number of syllables which are actually accented by stating :

(iv) Do not assign an accent to a syllable if a 'nearby' syllable within the same Intonation Unit is already accented.

The term 'nearby' is deliberately vague and can be made more specific in a number of ways, each of which would result in a different set of accent patterns. To simplify the discussion we shall assume here that 'nearby' is interpreted as 'adjacent' which results in a fluent and acceptable set of accent patterns. It should however be remembered that 'nearby' might equally well be interpreted as 'less than x syllables away' or even as 'less than x milliseconds away' where x has some integer value.

To take a few examples :

(9)a. Elle parle. (she speaks)
(9)b. Elle parlait. (she was speaking)
(9)c. Elle parlait français. (she was speaking French)
(9)d. Elle savait très bien parler le français. (she could speak French very well)
(9)e. Elle ne savait pas très bien parler le français. (she couldn't speak French very well).

The words 'pas' 'très' and 'bien' are all accentable in French despite the fact that they are function words. The rules given above generate the following patterns :

(10)a. Elle 'parle.
(10)b. Elle par'lait.
(10)c. Elle 'parlait fran'çais.
(10)d. Elle 'savait 'très bien 'parler le fran'çais.
(10)e. Elle ne 'savait 'pas très 'bien par'ler le fran'çais.

The fact that (ii) is ordered before (iii) ensures that the final syllable of each Intonation Unit will retain its accent. There is in fact independent evidence that the accent assigned to the final syllable of an Intonation Unit does not follow the same rules as the word-initial and word-final accent found elsewhere. Clitic syllables like "le", "en", "vous" are normally unaccentable as can be seen in (11a, 11b) :

(11)a. Je le tra'duis. (I translate it)
(11)b. Vous vous en a'llez. (you are going away)

When these syllables occur in final position, however, they are assigned prominence as in (12a-12c).

(12)a. 'Traduis-'le! (translate it!)
(12)b. 'Allez-vous 'en! (Go away!)
(12)c. 'Où allez-'vous?. (Where are you going?)

We have accounted for unemphatic intonation patterns in English and French (Hirst 1983, Hirst & Di Cristo 1984) by assuming that High and Low tones are attached directly to the accent group (which for this reason we have preferred to call the Tonal Unit) as well as to the higher order Intonation Unit in accordance with a tonal template with the following form:

(13)a. French : Tonal Unit (L H) Intonation Unit (L (H L))

(13)b. English : Tonal Unit (H L) Intonation Unit (L (H L))

This results in a non-linear prosodic structure which is submitted to linearisation constraints, projecting the tonal segments onto a single tonal tier.

A rule of downstepping also applies under slightly different conditions in English and French. This rule, which can be roughly formulated as follows :

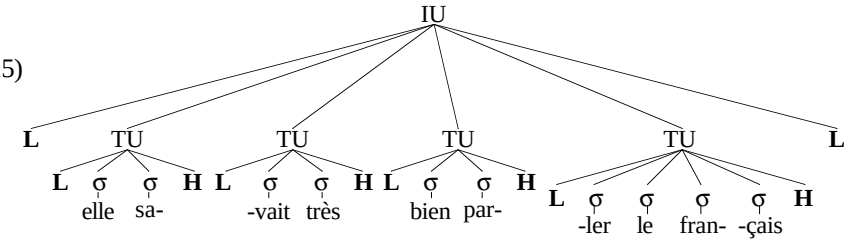
(14) Downstep : H L H → H ∅ D

converts a sequence HLH into a sequence H D (where D represents a downstepped (lowered) high tone[13]. The rule appears to be style/dialect dependent in English: systematic downstepping is said to be more frequent in Standard British English (RP) than in either Scottish English or American English (Hirst in press). In French, the downstepping rule seems dependent on the syntactic mode of the sentence, applying systematically in both yes/no-questions and wh-questions but only to the final tonal unit in assertions (Hirst & Di Cristo 1984, Di Cristo in press).

The choice of the final tone in the Intonation Unit seems to depend on pragmatic and semantic constraints in both English and French. In English and French a high final tone (final rise) is found both in continuatives and in questions and is generally held to be more common in Yes/No questions than in WH-questions in both languages.

Applied to sentence (9d) above, assuming the stress-pattern assigned in (10d), and assuming that a terminal falling pitch is chosen, templates (13a) will generate the following non-linear prosodic structure (where the symbol σ stands for 'syllable') :

(15)

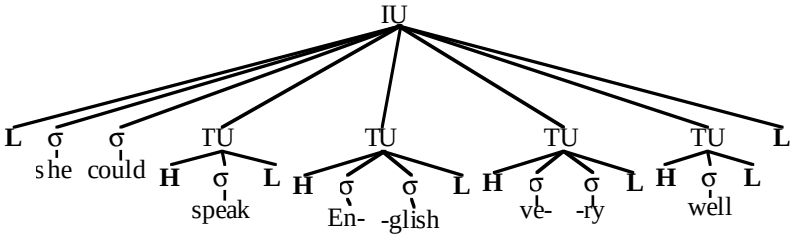


whereas the English equivalent :

(16) She could 'speak 'English 'very 'well.

would be assigned the structure :

(17)



After application of Downstep and other rules of surface phonological interpretation these non-linear structures will be converted to the bi-linear sequence :

(18)a.	elle sa-	-vait très	bien par-	-ler le français
	M T	L H	L H	D B

(18)b.	she could	speak	English	very	well
	M	T	D	D	D B

We have argued elsewhere (Hirst 1994; Hirst and Di Cristo in press) that representations such as these can subsequently be converted into phonetic reresentations which in turn can be used to generate synthetic fundamental frequency curves. Representations of this type have been used in two recent applications to speech synthesis for French, one based on a stochastic model (Véronis et al. 1997) and another based on a metrical model of French rhythm and intonation (Di Cristo et al. 1997) with very encouraging results.

4. Conclusion

The automatic generation of large quantities of symbolic data will without doubt make it possible in the next decade to develop a far more complete description of the prosody of different languages than is at present feasible. It remains to be seen how far the description of other languages, drawing both on existing work and on our own future invesigations, can be accomodated within a model such as we have outlined in this chapter. It appears to us quite probable that a number of characteristics of the phonological models which are current today, (including our own), will not stand up to empirical testing by such large quantities of data. It is also quite possible that a number of characteristics of the different modules which we have outlined above will need to be improved in order to account for the observed data. It is, however, our view that a model of intonation such as that which we have sketched above, specifying explictly different levels of representation, will continue to be an extremely interesting tool for comparing the prosodic systems of different languages.

CNRS ESA 6057, Laboratoire Parole et Langage
Institut de Phonétique, Université de Provence
13621 Aix en Provence, France
{Daniel.Hirst; Albert.DiCristo; Robert.Espesser}@lpl.univ-aix.fr

---

Notes

[1]In particular, our thanks for many helpful comments and remarks to participants at a number of different meetings incuding the 1st International ESCA conference on speech synthesis in Autrans (September 1990) the semi-plenary session on prosodic models at ICPhS XII in Aix en Provence (August 1991), the Prosody Seminar at La-Baume-les-Aix (May 1992), the ESCA Workshop on Prosody in Lund (September 1993), the National Language Institute 2nd International Symposium in Tokyo (June 1994), the AAAI-94 Workshop on the Integration of Speech and Natural Language Processing in Seattle (August 1994), the International Symposium on Prosody in Yokohama (September 1994), the IEEE/ESCA meeting on speech synthesis at New York (September 1994), the COST 233 meeting in Lausanne (November 1994), the Japic First Research Meeting at the National Language Institute in Tokyo (February 1995), the HCM Workshop on Dialog and Discourse Prosody in Stuttgart (February 1995) and the 21st Journées d'Etude sur la Parole in Avignon (June 1996), where different parts of this study were presented.

[2]These "target points" are essentially the "turning points" described by Gårding 1977, 1983, except that, unlike turning points, target points do not necessarily imply a change of direction: downstepped and upstepped targets can thus be generated directly as described below.

[3] Experimental evidence (Cavé, Hirst & Rossi 1986) suggests that the exact localisation of this inflection point is not crucial for the quality of synthesised speech. In natural speech a certain assymmetry is in fact observed with the inflection point being closer to the higher of the two targets.

[4] The "Rise/Fall/Connection" model described by Taylor (1994) is in fact formally identical to a quadratic spline function except that it also allows the possibility of linear transitions as well as curvi-linear transitions between target points with different F0 values. We have seen no convincing evidence that this option is ever in fact necessary. Note that curvi-linear transitions between successive targets with the same F0 value will be identical to linear transitions so that no special option is in fact needed to generate these with a quadratic spline function.

[5] The closest approximation to a quadratic spline function is given by using horizontal plateaux of approximately 1/3 of the duration between two target points.

[6] A similar result has now been obtained for the German recordings of Eurom1.
[7] When *a* is greater than 1 the re-iterated function diverges; when it is less than 0 it oscillates.
[8]The case of D and U is somewhat different. Phonological arguments (Clements & Ford 1979) suggest that D is equivalent to a sequence LH from which it generally seems to be derived diachronically. If this is taken literally, we should expect the U and D targets to be closer to the middle of the ptich range than those for T or H and B or L respectively. The data analysed here is not consistent with regard to this prediction, perhaps due to the fairly small number of examples of these tones.
[9] One attempt to provide a system of notation for English intonation based on purely formal criteria was that of Hirst (1977) who proposed a representational system based on five abstract features : [±stress; ±centre; ±emphasis; ±boundary ±terminal].
[10] More recently, Halle & Idsardi (1996) have claimed that grouping is in fact linguistically more basic than prominence which should be considered "a by-product of the grouping of the elements into constituents." (p 439)
[11] Studies of durational effects, in particular Reis (1995) for Brazilian Portuguese, Le Besnerais (1996) for Spanish, suggest at least the possibility that the left/right headedness of accent groups might be a parameter which distinguishes Germanic languages from Romance languages in general (Hirst et al 1993; Hirst & Di Cristo in press).
[12] More detailed accounts of the intonation systems of the two languages within the same basic framework are to be found in Di Cristo (in press) and Hirst (in press).
[13] In line with recent tendancies in phonology this downstep "rule" could be reformulated as a principle stating that under certain conditions an underlying L tone is delinked from the Tonal Unit, i.e. it no longer operates as a target but still has the effect of lowering the following H tone which is consequentlyinterpreted as a surface D tone.

References

d'Allessandro, C. and Mertens, P. 1995. Automatic pitch-contour stylisation using a model of tonal perception. *Computer Speech and Language* 9, 257-288.
Bagshaw, P. 1994. *Automatic Prosodic Analysis for Computer-Aided Pronunciation Teaching.* PhD thesis, University of Edinburgh.
Bosch (ten), 1993. On the automatic classification of pitch movements. in *Proceedings of the 3rd European Conference on Speech Communication and Technology* (Berlin 1993), 781-784.
Bruce, G. 1985. Structure and function of prosody. in Guérin, B. & Carré, R. (eds) *Franco-Swedish Seminar on Speech* (Grenoble) 2, 549-559.
Bruce, G. 1989. Report from the IPA working group on suprasegmental categories.*Working Papers* 35 (Lund University), 25-40
Campbell, W.N., 1992. *Multi-level Timing in Speech,* PhD Thesis, University of Sussex.
Campbell, W.N., 1992. Multi-level Timing in Speech, this volume, 000-000.
Campione, E., Flachaire, E., Hirst, D.J. and Véronis, J. 1997. Stylisation and symbolic coding of F0, a quantitative model. in *Intonation : Theory, Models and Applications. Proceedingsof an ESCA Workshop. (*Athens, 1997), 71-74.
Cavé, C.; Hirst, D.J. & Rossi, M. 1986. Pitch of glissandos in speech sounds. *Proceedings of the 12th International Congress on Acoustics* (Toronto, 1986)
Chan, D., Fourcin, A., Gibbon, D., Granstrom, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Tracoso, I., Veld, C. & Zeiliger, J. 1995. EUROM - a spoken language resource for the EU. in *Proceedings of the 4th European Conference on Speech Communication and Technology* (Madrid 1995), 000-000.
Chomsky, N. 1964. *Current Issues in Linguistic Theory* (Mouton; La Haye)
Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.
Clements, G.N & Ford, K.C. 1979. Kikuyu tone shift and its synchronic consequences. *Linguistic Inquiry* 10 (1), 179-210.
Cruttenden, A. 1986. *Intonation*. Cambridge University Press, Cambridge.
Di Cristo, A. 1985. *De la microprosodie à l'intonosyntaxe*. Publications de l'Université de Provence, Aix en Provence.
Di Cristo, A. in press. Intonation in French. in Hirst & Di Cristo (eds) in press. 196-219.
Di Cristo, A., Di Cristo, P. and Véronis, J. 1997. A metrical model of rhythm and intonation for French text-to-speech synthesis. in *Intonation : Theory, Models and Applications. Proceedingsof an ESCA Workshop. (*Athens, September 1997), 83-86.
Di Cristo, A. & Hirst, D.J. 1986. Modelling French micromelody : analysis and synthesis. in Kohler (ed.) 1986) *Prosodic Cues for Segments* (= *Phonetica* 43 (1-3)), 11-30.
Di Cristo, A. & Hirst, D.J. 1997. L'accentuation non-emphatique en français : stratégies et paramètres. in J. Perrot (ed) 1995. *Polyphonie pour Ivan Fónagy*. Paris, Harmattan 71-101.

Espesser, R. 1982. Un système de détection du voisement et de F0. *Travaux de l'Institut de Phonétique d'Aix* 8, 241-261
Fant, G., Kruckenberg, A. & Nord, L. 1991. Language specific patterns of prosodic and segmental structures in English, Swedish and French. *Proceedings of the 12th International Congress of Phonetic Sciences*, 118-121.
Fónagy, I. 1980. L'accent en français: accent probabilitaire. *Studia Phonetica* 15, 123-233.
Fujisaki, H. & Nagashima, S. 1969. A model for synthesis of pitch contours of connected speech. *Annual Report Engineering Research Institute : University of Tokyo*. 28, 53-60.
Fujisaki, H. 1988. A note on the phyiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. in O. Fujumura (ed) *Vocal Physiology. Voice Production, Mechanisms and Findings.* Raven Press 347-355.
Fujisaki, H. 1997. Prosody, models and spontaneous speech. in Sagisaka et al. (eds.) 27-42.
Gårding, E. 1977. The importance of turning-points for the pitch patterns of Swedish accents." in L.M.Hyman (ed.) *Studies in Stress and Accent (= Southern California Occasional Papers in Linguistics 4)*, 27-36
Gårding, E. 1983. A generative model of intonation. in Cutler & Ladd (eds) *Intonation : Models and Measurements*. (Springer, Berlin), 11-26
Geoffroi, E. 1993. A pitch contour analysis guided by prosodic event detection. in *Proceedings of the 3rd European Conference on Speech Communication and Technology* (Berlin 1995), 793-796.
Halle, M. & Vergaud, J.R. 1989. *An Essay on Stress*. MIT Press, Cambridge Mass.
Halle, M. & Idsardi, W. 1996.
Hamon, Moulines, E. & Charpentier 1989. A diphone system based on time domain modifications of speech. *Proc. Int. Conf. Assp.* , 239-241.
Hart, ('t), J.; & Collier, R. 1975. Integrating different levels of intonation analysis. *Journal of Phonetics* 3, 235-255.
Hart, J. ('t) 1991. F0 stylisation in speech : straight lines versus parabolas. *J. Acoust. Soc. Am*. 6, 3368-3370.
Hart, J. ('t), Cohen & Collier, R. 1990. *A Perceptual Study of Intonation : an Experimental-Phonetic Approach to Speech Melody*. Cambridge University Press, Cambridge.
Hermes, D.J. & Van Gestel, J.C. 1991. The frequency scale of speech intonation., *J. Ac. Soc. of Am*., vol. 90, pp. 97-102.
Hirst, D.J. 1977. *Intonative Feautres : a Syntactic Approcah to English Intonation.* Mouton, The Hague.
Hirst, D.J. 1980. Un modèle de production de l'intonation. *Travaux de l'Institut de Phonétique d'Aix* 7, 297-315
Hirst, D.J. 1983. Structures and categories in prosodic representations. in Cutler & Ladd 1983) *Prosody : Models & Measurements* (Springer, Berlin) , 93-109
Hirst, D.J. 1987. *La description linguistique des systèmes prosodiques : une approche cognitive* Thèse de Doctorat d'Etat, Université de Provence
Hirst, D.J. 1992. Prediction of prosody : an overview. in G.Bailly & C.Benoît (eds) *Talking Machines : Theories, Models and Applications*. (Elsevier Science Publishers)
Hirst, D.J. 1993. Detaching intonation phrases from syntactic structure. *Linguistic Inquiry* 24 (4), 781-788.
Hirst, D.J. (in press). Intonation in British English. in Hirst & Di Cristo (eds) in press. 56-77.
Hirst , D.J. & Di Cristo, A. 1984. French intonation : a parametric approach. Die Neueren Sprachen 83 (5), 564-569.
Hirst , D.J. & Di Cristo, A. in press. A survey of intonation systems. in Hirst & Di Cristo (eds) in press. 1-44.
Hirst , D.J. & Di Cristo, A. (eds) in press. *Intonation Systems : a Survey of Twenty Languages.* Cambridge University Press; Cambridge.
Hirst, D.J. & Espesser, R. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix* 15, 71-85
Hirst, D.J.; Di Cristo, A.; Le Besnerais, M.; Najim, Z. & Nicolas, P. 1993. Multi-lingual modelling of intonation patterns. *Proceedings ESCA Workshop on Prosody*. (Lund, September 1993), 204-207
Hirst, D.J.; Nicolas, P. & Espesser, R. 1991. Coding the F0 of a continuous text in French : an Experimental Approach. *12° Congrès International des Sciences Phonétiques* (Aix en provence), Vol. 5, 234-237.
Holmes, V. 1996. A crosslinguistic comparison of the production of utterances in discourse. *Cognition* 54: 169-207.
House. D. 1990. *Tonal Perception of Speech. (= Travaux de l'Institut de Linguistique de Lund 24)* Lund, Lund University Press.

Ladd, D.R. 1993. Notes on the phonology of prominence. *Working Papers (University of Lund)* 41, 10-15.
Le Besnerais, M. 1995. *Parámetros rítmicos para el estudio contrastado del francés y del espaõl contemporáneos.* Doctoral thesis, University of Barcelona.
Martin, P. 1981. Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne. *Actes des 12e Journées d'Etudes sur la Parole* (Montréal), 221-232.
Morlec, Y., Bailly, G. & Aubergé, V. 1995. Synthesis and evalutation of intonation with a superposition model. *Proceedings of the 4th European Conference on Speech Communication and Technology* (Madrid 1995), 2043-2046.
Nicolas, P. 1989. *Amplitude des variations micromélodiques des obstruantes voisées en fonction de la hauteur de la voix*. DEA Dissertation, Université de Provence.
Nicolas, P. 1995. *La structuration prosodique du texte lu en français.* Doctoral thesis, Université de Provence.
Nicolas, P. & Hirst, D.J. 1995. Symbolic coding of higher level characteristics of fundamental frequency curves. *Proceedings of the 4th European Conference on Speech Communication and Technology* (Madrid 1995)
Pasdeloup, V. 1990. *Modèle de règles rythmiques du français appliqué à la synthèse de la parole.* Doctoral thesis, Université de Provence.
Pierrehumbert, J. this volume. Tonal elements and their alignment.
Pierrehumbert, J. & Beckman, M. 1988. *Japanese Tone Structure*. MIT Press. Cambridge, Mass.
Reis, C. 1995. *L'interaction entre accent intonation et rythme en portugais brésilien*. Doctoral thesis, Université de Provence.
Roucos & Wilgus 1985. High quality time-scale modification for speech. *Proceedings of ICASSP*, 493-496.
Rossi, M., Di Cristo, A., Hirst, D.J., Martin, P. & Nishinuma, Y. 1980. *L'intonation : de l'acoustique à la sémantique.* Paris, Klincksieck.
Sagisaka, Y., Campbell, N., Higuchi, N. (eds.) 1997. *Computing Prosody. Computational Models for Processing Spontaneous Speech*. New York, Springer.
Saussure, F. (de) 1916. *Cours de Linguistique Générale.* Paris, Payot.
Scheffers, M.T.M. 1988. Automatic stylization of F0-contours. in *Proceedings of 7th FASE symposium : Speech '88* (Edinburgh),
Selkirk, E. 1978. *On prosodic structre and its relation to syntactic structure.* Bloomington, Indiana University Press.
Selkirk, E. 1984. *Phonology and syntax: the relation between sound and structure.* Cambridge, Mass., MIT Press.
Selkirk, E. 1986. On derived domains in sentence phonology. *Phonology Yearbook* 3, 371-405.
Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. ToBI : a standard for labelling English prosody. *Proc. Internal. Conf. Spoken Language Processing* . 2, 867-870.
Strangert, E. & Aasa, A. 1996. Strangert, E. and Aasa, A. 1996. Evaluation of Swedish prosody within the MULTEXT-SW project. *TMH-QPSR* 2/1996 *(Speech, Music and Hearing - Quarterly Progress and Status Report)*, KTH, Stockholm, Sweden, 37-40.
Taylor, Paul 199. Automatic recognition of intonation from F0 contours using the Rise/Fall/Connection model *Proceedings Eurospeech '93*. (Madrid, September 1993), 789-792
Terken, J. 1991. Fundamental frequency and perceived prominenc of accented syllables. *J. Acoust. Soc. Am.* 89. 1768-1776.
Terken, J. & Hermes, D. this volume. The perception of prominence.
Trubetzkoy 1949. Grundzüge der Phonologie. (French translation by J. Cantineau 1957) Principes de phonologie. Klincksieck; Paris.
Véronis, J., Di Cristo, P., Courtois, F., Lagrue, B. 1997. A stochastic model of intonation for French text-t-speech synthesis. *Proceedings of the 5th European Conference on Speech Communication and Technology* (Rhodes 1995), 2643-2646.
Véronis, J.; Hirst, D.J.; Espesser, R. & Ide, N. 1994.NL and speech in MULTEXT. *Proceedings AAAI-94 Workshop of the Integration of Speech and Natural Language Processing.* (Seattle; August 1994).
Wenk & Wiolland, F. 1982. Is French really syllable-timed? *Journal of Phonetics* 10(2), 193-216.