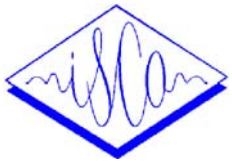


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# THE SYMBOLIC CODING OF SEGMENTAL DURATION AND TONAL ALIGNMENT: AN EXTENSION TO THE INTSINT SYSTEM.

*Daniel Hirst*
CNRS & Université de Provence. email: daniel.hirst@lpl.univ-aix.fr
http://www.lpl.univ-aix.fr/~hirst



## ABSTRACT

This paper presents work based on an analysis-by-synthesis approach which aims to develop a reversible coding system for prosody, capable of deriving a 'linguistic-like' surface phonological representation directly from acoustic data that is sufficient to reproduce a synthetic version of the original utterance without significant loss of linguistic information. With such a coding system, capable of representing any significant prosodic distinctions, the task of predicting such representations would be greatly simplified, becoming one of mapping between sets of symbolic representations. This approach has already been applied to the stylisation and symbolic coding of fundamental frequency curves by means of the INTSINT transcription system. An automatic version has also been proposed. This paper presents a preliminary proposal for an extension to the INTSINT system to cover segmental duration and the relative alignment of phonematic and tonal segments.

## 1. INTRODUCTION: INTSINT TRANSCRIPTIONS

**INTSINT** (an **IN**ternational **T**ranscription **S**ystem for **INT**onation) was developed during the preparation of a study of the intonation of twenty languages [11] and was used to transcribe examples of intonation patterns for nine of these: British English, Spanish, European Portuguese, Brazilian Portuguese, French, Romanian, Russian, Moroccan Arabic and Japanese. It aims to capture the surface distinctions used in different languages for building distinctive intonation patterns. Unlike many other transcription systems, including in particular ToBI [20], [19], INTSINT is entirely concerned with the representation of prosodic form rather than of prosodic function. In this sense it can be thought of as a prosodic equivalent of a narrow IPA transcription system for segmental transcriptions.

INTSINT represents an intonation pattern as a sequence of 'tones', coding the relative height of the significant "target points" of the pattern. Three of these tones: *Top*, *Mid* and *Bottom* are assumed to be defined globally with respect to the speaker's pitch range. Three other tones: *Higher*, *Same*, *Lower* are defined locally with respect to the preceding tone. Two further tones *Upstepped* and *Downstepped* are similar to *Higher* and *Lower* but imply a smaller interval with respect to the preceding tone. Typically, *Upstepped* and *Downstepped* are used in iterative sequences whereas *Higher* and *Lower* will generally correspond to peaks and valleys. Table 1 (from [14]) shows the orthographic and iconic symbols used in INTSINT.

Table 1: Orthographic and iconic symbols for the INTSINT coding system.

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

The choice of tonal symbols implies a quantification of the frequency domain. The alignment of the two sets of symbols, however, has not yet been the object of similar quantification. Instead, typically, the symbols are aligned graphically and analogically as in the following, a transcription of the French utterance "Il faut que je sois à Grenoble Samedi vers quinze heures." (I have to be in Grenoble by Saturday 3 p.m.):

(1) [ilfok@Z@swazagR@nObl][samdivERk~Ez9R]
[ ⇑ ↓ ↑ ↓ ⇑] [ ⇑ ↓ ↑ ⇓]

The fact that INTSINT codes prosodic form rather than prosodic function means that it can be used as a data-driven tool for automatically extracting 'linguistic-like' information from acoustic data. In conjunction with MOMEL, which provides an automatic stylisation of F0 curves [12], [13] INTSINT has been used as a reversible coding system for F0 curves for a number of languages [1], [15], [21]. and the representation system developed has now been implemented in two text-to-speech systems for French [4], [22],

In the rest of this paper, rather than the iconic symbols used in (1), I use the orthographic symbols (T, M, B etc.) which (with the exception of D and L) have been integrated into the SAM phonetic alphabet SAMPA[23]. This alphabet is particularly suitable for computer-coding since it makes use only of symbols in the ASCII/ANSI range 32 to 126 and can consequently be used without problem for transferring transcription files between computers using different systems, where higher ASCII number characters are usually not compatible. For a recent proposal to extend SAMPA (X-SAMPA) to cover all current IPA symbols, see [24].

## 2. SURFACE REPRESENTATIONS OF SEGMENTAL DURATION AND TONAL ALIGNMENT.

INTSINT does allow for some degree of tonal alignment by the use of square brackets as in (1), to indicate points of synchronisation. Here I explore some possible extensions to the INTSINT system which would provide for a completely symbolic transcription of an utterance as a linear sequence of symbols including features of duration (lengthening etc) and tonal alignment. Such a notational system should ideally be as independent as possible of any particular prosodic theory. Indeed one of the aims of such a system would be to provide a

means to compare alternative hypotheses in a common notational form. Implemented in conjunction with a text-to-speech system it would then be possible to evaluate fairly directly the relative merits of different theories.

The INTSINT system can be extended to provide symbolic coding for segmental duration by coding durations of segments with the symbols [-, +] for short and long respectively Segments with no explicit duration symbol are assumed to be of average length. Extra degrees of lengthening or shortening can be indicated by repeating the duration symbol (a--, a++, a---,a+++ etc.) [1].

Two types of representations are possible: linear representations and tiered representations. In the latter, segments and tones are coded in different tiers (or streams, columns, fields…) and there is consequently no ambiguity between the two sets of symbols used. For linear representations, there is overlap between the segmental symbols and the tonal symbols: SAMPA D = IPA /ð/, T = /θ/ S = /ʃ/ etc. In order to distinguish tones and segments in linear representations, I follow Wells' proposal [24] (following a suggestion by Dafydd Gibbon) to include tonal symbols in angled brackets < > acting as 'tier escape' symbols.

For the relative alignment of tones and segments, different levels of representation are assumed [14]. At a surface level the simplest solution is to suppose that tones are aligned with respect to phonematic segments. The relative timing of the tone to the preceding segment can be specified by means of 4 symbols '[', '+',' -', and ']' corresponding to *beginning*, *early*, *late* and *end* respectively. When there is no diacritic it is assumed that the tone symbol is aligned with the *middle* of the preceding segment.

Example (2) illustrates the utterance "Like this" with a specific alignment of the pitch pattern <MTDB> All the segments are of average length except the final vowel, which is long, and the final consonant, which is extra-long.

| | |
|---|---|
| _ | M |
| l | |
| AI | T+ |
| k | |
| D | |
| I+ | D[ |
| s++ | |
| _ | B |

(2). A tiered INTSINT representation of a reading of the utterance "Like this".

[1] This is a slight divergence from SAMPA which proposes to use '-' as a separator. The symbol '.' (not used in X-SAMPA) could however be used for this purpose. It seems preferable to keep paired symbols such as '+' and '-" for paired interpretations such as that proposed here. Note that the colon is used in IPA to represent distinctive phonological length, not the actual physical duration of the segment. Thus /i:-/ would represent a shorter than average /i:/ segment.

Using a linear transcription the same utterance would be coded:

_<M>lAI<T+>kDI+<D[>s++_<B>

(3). A linear INTSINT representation of example (2)

To provide auditory assessment, the system has been interfaced with the MBROLA diphone speech synthesiser [6], [7] by a MacPerl script *int2pho* which takes as input a tiered INTSINT file (.int) (like that of (2) above) and provides as output an input file for MBROLA (.pho) with appropriate durations and pitch values. With the MBROLA synthesiser, the appropriate set of diphones and a table of mean durations for the phonemes of the language), *int2pho* can be used for any of the languages currently available for the MBROLA project.

The script calculates segmental durations and pitch targets using a number of parameters. Those for duration (with default values in parentheses) are: *tempo(1)*, *extrashort(0.5)*, *short(0.75)*, *long(1.5)*, *extralong(2)*. Parameters for F0 are *key(90)*, *range(2)*, *lower(0.5)*, *higher(0.5)*, *upstep(0.25)*, *downstep(0.25)*, *same(0.05)*. Parameters for alignment are *beginning(0)*, *early(25)*, *middle(50)*, *late(75)*, *end(100)*. All the parameters can be modified by the user in his transcription. Thus for example tempo is assigned the default value of 1, but a line containing :

<parameter tempo=1.2>

will increase the duration of segments from there on by 20%.

Segmental duration is established by looking up the mean value for the phoneme, multiplying this by the appropriate lengthening or shortening parameter and multiplying the result by the tempo parameter.

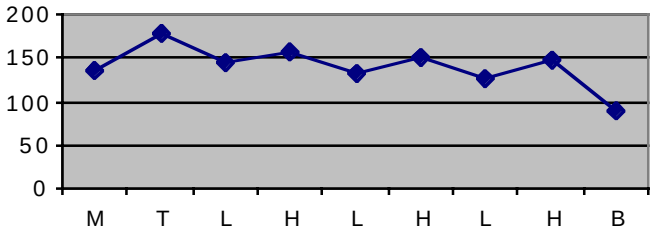
An F0 target point Pi is calculated by the following algorithm(from [14]):

*bottom* = *key*; *top* = *key***range*; *mid* = (*top* + *bottom*)/2

| | |
|---|---|
| <B>: | $P_i$ = *bottom*. |
| <T>: | $P_i$ = *top* |
| <M>: | $P_i$ = *mid*. |
| <H> : | $P_i = P_{I-1} + (top - P_{I-1})$**higher* |
| <U> : | $P_i = P_{I-1} + (top - P_{I-1})$**upstep* |
| <L> : | $P_i = P_{I-1} - (P_{I-1} - bottom)$**lower* |
| <D> : | $P_i = P_{I-1} - (P_{I-1} - bottom)$**downstep* |
| <S> : | $P_i = P_{I-1} - (P_{I-1} - bottom)$**same* |

(4) Algorithm for the interpretation of INTSINT targets

An interesting side-effect of (4) is that a sequence of <D> tones will asymptotically downstep (as observed by [18]) and that a sequence of alternating H and L values will asymptotically downdrift as shown in (5) without any need for a specific declination component.




(5) Sample output of (4) showing asymptotic downdrifting.

By default, the algorithm (4) is applied on a linear scale. The same algorithm can be applied on a log scale by modifying the parameter *scale* (i.e. by including the line <parameter scale=log>). Other scales (Mel, Bark, ERB etc.) could easily be provided.

The relative alignment of the tone with respect to the preceding phoneme is determined by the presence or absence of an alignment diacritic. When there is none, the tone is aligned by default with the parameter *middle*, initialised to 50%. The other parameters determined by the diacritics [, -, +, ] are initialised respectively to 0, 25, 75 and 100%.

MBROLA in its present form interpolates linearly between target points. INTSINT assumes in fact that between target points there is a curvilinear (quadratic) interpolation, which enables a much sparser representation of the intonation pattern. It is hoped that a future version of MBROLA will allow users the option of interpolation with a quadratic spline function as described by [8], [14].

## 3 HIGHER-LEVEL PROSODIC CONSTITUENTS.

The aim of this work is to provide a framework in which the adequacy of different models of prosody can be evaluated. Most models of prosody, in fact, do not assume that tones are aligned with phonemes. An exception is recent work by Arvaniti et al. [1] who claim that pre-nuclear rising accents in Greek are best analysed as a sequence of tones L and H which are aligned respectively with the onset of the stressed syllable and the vowel onset of the post-accentual syllable. In standard INTSINT notation this would be represented as in the following example:

(6) …pa'ranoma…
L H

In the extended INTSINT notation I am proposing here this could be represented:

(7) …pa<L]>'ran<H]>oma

More often, some higher-level prosodic constituent is taken to be the level at which tones and segments are linked.

In order to specify for example that a tone is aligned in some way with a syllable or a foot or a syllable rime rather with a particular segment it is necessary to indicate the appropriate sequence. This can be done by a tiered representation such as:

_
nV M T
TIN BU
_

(8). A tiered INTSINT representation of the word "nothing" aligned syllable by syllable with the sequences <MT> and <BU>.

For linear representations, I propose to adopt a notation called *polymetrical expressions* recently proposed for the representation of simultaneous sequences in music [2]. A polymetrical expression (A, B, C…) is taken to represent the simultaneous realisation of the sequences A, B, C etc.. This notation has the advantage that it could be extended to cover any number of different tiers rather than just two. Example (8) using this type of representation would be coded:

_ (nV, <MH>)(TIN, <BU>) _

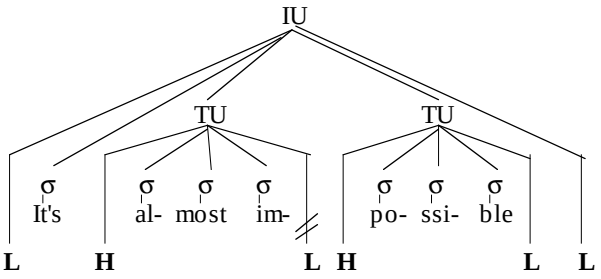
(9). A polymetrical INTSINT representation of the word "nothing" aligned syllable by syllable with the sequences<MT><BU>.

A somewhat different way of presenting the example from Greek quoted above would be to claim that the individual tones L and H are not aligned individually with the corresponding phonemes but rather that the sequence <LH> is aligned with the beginning and end of the sequence /ran/.

With the type of representation proposed above, the alignment of the sequence <LH> with the sequence /'ran/ would be represented;

(10) …pa('ran, <L[H]>)oma…

To illustrate a more complex example, [9] proposes a multi-linear underlying representation of an intonation pattern of the form:




(11) A multi-linear representation of the intonation of the utterance "It's almost impossible" in British English.

where tones are linked to different higher level prosodic constituents. The double bar across the link to the second L tone is taken to represent the fact that this tones is 'floating', that is that it does not surface as a tonal target although it will influence the pitch value of the subsequent H tone.

With the transcription system proposed here this would be represented:

(12) (Its(O:lm@UstIm, <HL>)(pQsIbl=, <H//L>), <LL>)

where the symbol '//' is taken as indicating that the following L tone is "floating".

The notation system described here provides a framework in which any multilinear prosodic representation could be described, and where the relative alignment of tonal segments to the different prosodic constituents is indicated using the same diacritics and parameters as that of surface representations described above.

## 4. DERIVING 'LINGUISTIC-LIKE' INFORMATION FROM ACOUSTIC DATA.

The next step in this project will be to develop a reverse model deriving a 'linguistic-like' representation from acoustic data.

For surface representations this is relatively simple[2]. Preliminary results from the application of this system to the Eurom1 corpus [4] will be presented.

For more abstract representations a number of challenging and interesting problems of phonetic interpretation arise which will be addressed in future work.

## REFERENCES


[1] Arvaniti, Amalia Ladd, D. R. and Mennen, Ineke 1998. Stability of Tonal Alignment: the case of Greek Prenuclear Accents. *Journal of Phonetics*

[2] Astésano, C.; Espesser, R.; Hirst, D.J. & Llisterri, J. 1997. Stylisation automatique de la fréquence fondamentale : une évaluation multilingue. *Proceedings 4th Congrès Français d'Acoustique*, 14-18 avril 1997, Marseille 441-443.

[3] Bel, B. 1992. Symbolic and sonic representations of sound-object structures. In M. Balaban, K. Ebcioglu, and O. Laske (eds.) *Understanding Music With AI.* Menlo Park: AAAI Press, pp.64-109.

[4] Chan, D. et al. 1995. Eurom - a spoken language resource for the EU. *Proceedings Eurospeech '95.* (Madrid) 867-870.

[5] Di Cristo, A., Di Cristo, P., Véronis, J. 1997. A metrical model of rhythm and intonation for French text-to-speech synthesis. *Proc. ESCA Workshop on Intonation*, Athens Sep. 1997, 83-86.

[6] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van Der Vrecken, O. 1996. The MBROLA project. Towards a set of high-quality speech synthesisers free of use for non-commercial purposes. *Proceedings ICSLP '96* (Philadelphia) 3: 1393-1396.

[7] Dutoit, T. 1997. *An Introduction to Text-to-Speech Synthesis*, Dordrecht: Kluwer.

[8] Hirst, D.J. 1983. Structures and categories in prosodic representations. in A. Cutler & D.R.Ladd (eds.) *Prosody: Models and Measurements.* Springer, Berlin. 93-109.

[9] Hirst, D.J. 1998. Intonation in British English. in Hirst & Di Cristo eds. 1998, 56-77.

[10] Hirst, D.J., Di Cristo, A. 1998.. A survey of intonation systems. in Hirst & Di Cristo eds. 1998, 1-44.

[11] Hirst, D.J., Di Cristo, A. (eds) *Intonation System. A Survey of Twenty Languages*. Cambridge; Cambridge University Press.

[12] Hirst, D., Di Cristo, A., Espesser, R. in press.. Levels of representation and levels of analysis for the description of intonation systems. In Horne, M. (ed.) in press.

[13] Hirst, D., Espesser, R. 1993. Automatic Modelling of Fundamental Frequency using a quadratic spline function, *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 15, 75-85.

[14] Hirst, D.J., Di Cristo, A & Espesser, R. in press. Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (ed.) in press

[15] Hirst, D.J.; A. Di Cristo, M. Le Besnerais, Z. Najim, P. Nicolas, P. Roméas 1993. Multi-lingual modelling of intonation patterns. *Proceedings ESCA Workshop on Prosody*. Lund, Septembre 1993, 204-207.

[16] Horne, M. (ed.) in press.. *Prosody: Theory and Experiment*, Dordrecht; Kluwer Academic Publishers.

[17] Mora, E. Hirst, D.J., Di Cristo, A. 1997. Intonation features as a form of dialectal distiinction. in *Proc. ESCA Workshop on Intonation*, Athens Sep. 1997, 247-250.

[18] Liberman, M., Pierrehumbert, J.,. 1984. Intonational invariance under changes in pitch range and length. in M. Aranoff and R. Oerhle (eds.) *Language Sound Structure: Studies in Phonology Presented to Morris Halle.* Cambridge, Mass.; MIT Press., 157-233.

[19] Pierrehumbert, J. in press. Tonal elements and their alignment. in M. Horne (ed.) in press.

[20] 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 ICSLP'92*, 2, 867-870, Banff, Canada.

[21] 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.

[22] Véronis, J., Di Cristo, P., Courtois, F. & Lagrue, B. 1997. A stochastic model of intonation for text-to-speech synthesis. *Proceedings Eurospeech '97* (Rhodes) 5: 2643-2646.

[23] Wells, J.C., Barry, W., Grice, M., Fourcin, A., Gibbon, D. .1992. Standard computer-compatible transcription. *Esprit project 2589 (SAM), Doc. no. SAM-UCL-037*. London, Phonetics and Linguistics Department, UCL

[24] Wells, J.C. 1995. Computer-coding the IPA: a proposed extension of SAMPA. ms.

http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm.


[2] The MacPerl script *int2pho* together with a preliminary version of the inverse script *pho2int* as well as example labelled files from the Eurom1 corpus [3] will be found at the address http://www.lpl.univ-aix.fr/~hirst