A framework for the multi-lingual analysis by synthesis of speech melody .

Daniel Hirst
CNRS Laboratoire Parole et Langage,
Université de Provence, Aix en Provence, France
daniel.hirst@lpl.univ-aix.fr

1 Introduction

This paper presents a framework for the multilingual analysis by synthesis of speech prosody which has been developed over a number of years in Aix-en-Provence and is currently being applied to and evaluated with a number of indo-european and non indo-european languages.

The analysis is based on the assumption that in all languages a certain number of prosodic functions are expressed by prosodic forms. While there is evidence that both the functions and forms of prosody present quasi-universal characteristics, the mapping between the two levels is clearly highly language (and dialect) specific.

The reversibility of the acoustic modelling and coding means that the only question the linguist need ask is whether the resynthesised speech expresses the same prosodic functions as the original signal: it is not necessary at this stage to identify or classify these functions, many of which have proved notoriously elusive to investigators.

Once a sufficient degree of approximation to the original message has been achieved by an appropriate symbolic annotation, it then becomes a task of relating these transcriptions to more abstract representations of prosodic functions, a task which it is hoped will prove amenable to techniques of natural language processing. It is further hoped that the fact that no a priori assumptions about the inventory of prosodic functions are incorporated into the transcription of prosodic forms will make it possible to evaluate the adequacy of different models of prosodic functions.

2 Analysis by synthesis of speech prosody

In recent work [16][11] it has been argued that the separation of form and function in the representation of speech prosody is a highly desirable condition for the analysis of the prosodic systems of natural languages. In the area of speech synthesis, by contrast, the representation of prosody often combines both form and function. In the project described here, the aim is to develop and implement a symbolic representation system for prosodic form without direct reference to prosodic function.

The symbolic representation system described can be derived automatically from acoustic data via a specification of the domains and units relevant for the analysis. The analysis is reversible so that the result of the symbolic coding of the data can be compared empirically with the original data in order to evaluate the appropriateness of the analysis. The specification of domains and units for each prosodic parameter thus becomes an explicit step in the modeling of the prosodic system in order to allow the user to implement and test different models of representation.

The prosodic parameters currently implemented in the ProZed framework, an ongoing implementation of the model, are segmental duration and fundamental frequency. But the same general framework could, and it is hoped will, be extended to include other parameters such as spectral tilt and voice quality. One specific characteristic of the implementation is that different domains and units can be specified for different parameters so that the units used to define the rhythm of an utterance, for example, are not necessarily the same as those used to define its pitch.

The system implements the symbolic representation of speech prosody as a set of hierarchical structures defining specific units for the interpretation of discrete symbols coding the absolute and relative pitch and duration of different units of speech.

The representation of rhythm has recently been described elsewhere [12] [13] while that of melody is presented in more detail below, but the two levels of representation have a certain number of characteristics in common. For both levels, a distinction is made between, on the one hand, local, short-term variability, (i.e. lexical or non-lexical discrete distinctions of tone and quantity) for which specific units are assumed, and on the other hand longer term variability involving higher order domains. Thus, for rhythm, it is assumed that within a specific rhythm domain, a constant *tempo* is defined which then serves as a default reference with respect to which shorter term variability is described.

In the same way, melody is described by means of melody domains within which the speaker's overall pitch reference level, referred to as his *key*, and the extent of variability, referred to as his *range*, are assumed to be constant.

2.1 Momel - a phonetic representation of pitch

The analysis of raw fundamental frequency curves for the study of intonation needs to take into account the fact that speakers are simultaneously producing an intonation pattern and a sequence of syllables made up of segmental phones. The actual raw fundamental frequency curves that can be analysed acoustically are the result of an interaction between these two components and this makes it difficult to compare intonation patterns when they are produced with different segmental material. Compare for example the intonation patterns on the utterances *Its for papa* and *It's for mama.*

The Momel algorithm attempts to solve this problem by factoring the raw curves into two components:

• a *macromelodic component* - modelled as a quadratic spline function.

This is assumed to correspond to the global pitch contour of the utterance, and which is independent of the nature of the constituent phonemes. The underlying hypothesis is that this macromelodic component is, unlike raw fundamental frequency curves, both continuous and smooth. It corresponds approximately to what we produce if we hum an utterance instead of speaking it.

- a *micromelodic component* consisting of deviations from the macromelodic curve - called a *micromelodic profile.*

This residual curve is assumed to be determined entirely by the segmental constituents of the utterance and to be independent of the macromelodic component.

The quadratic spline function used to model the macromelodic component is defined by a sequence of target points, (couples <s, Hz> each pair of which is linked by two monotonic parabolic curves with the spline knot occurring (by default) at the midway point between the two targets. The first derivative of the curve thus defined is zero at each target point and the two parabolas have the same value and same derivative at the spline knot. This in fact defines the most simple mathematical function for which the curves are both continuous and smooth.

On the one hand, the two French utterances "à ma maman" (to my mummy) and "à ton papa!" (to your daddy) could thus be modelled with the same target points (hence the same macromelodic component). See Figure 1.

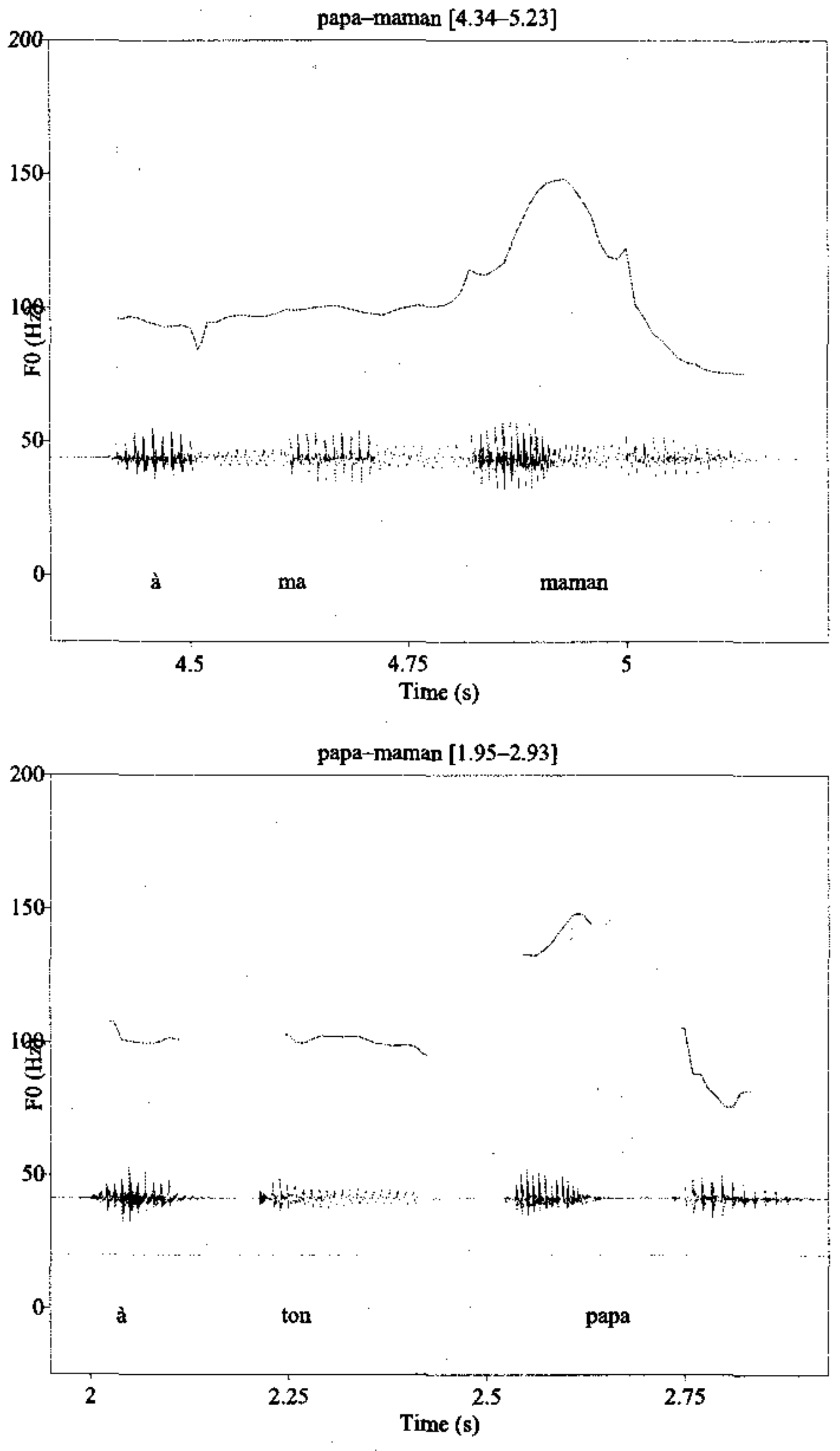

Figure 1. Fundamental frequency and wave-form for the two French utterances "A ma maman!" and "A ton papa!"

The same words pronounced as a question "à ma maman?" and "à ton papa?" would also have the same target points but which would probably be different from those of the first pair. See figure 2.

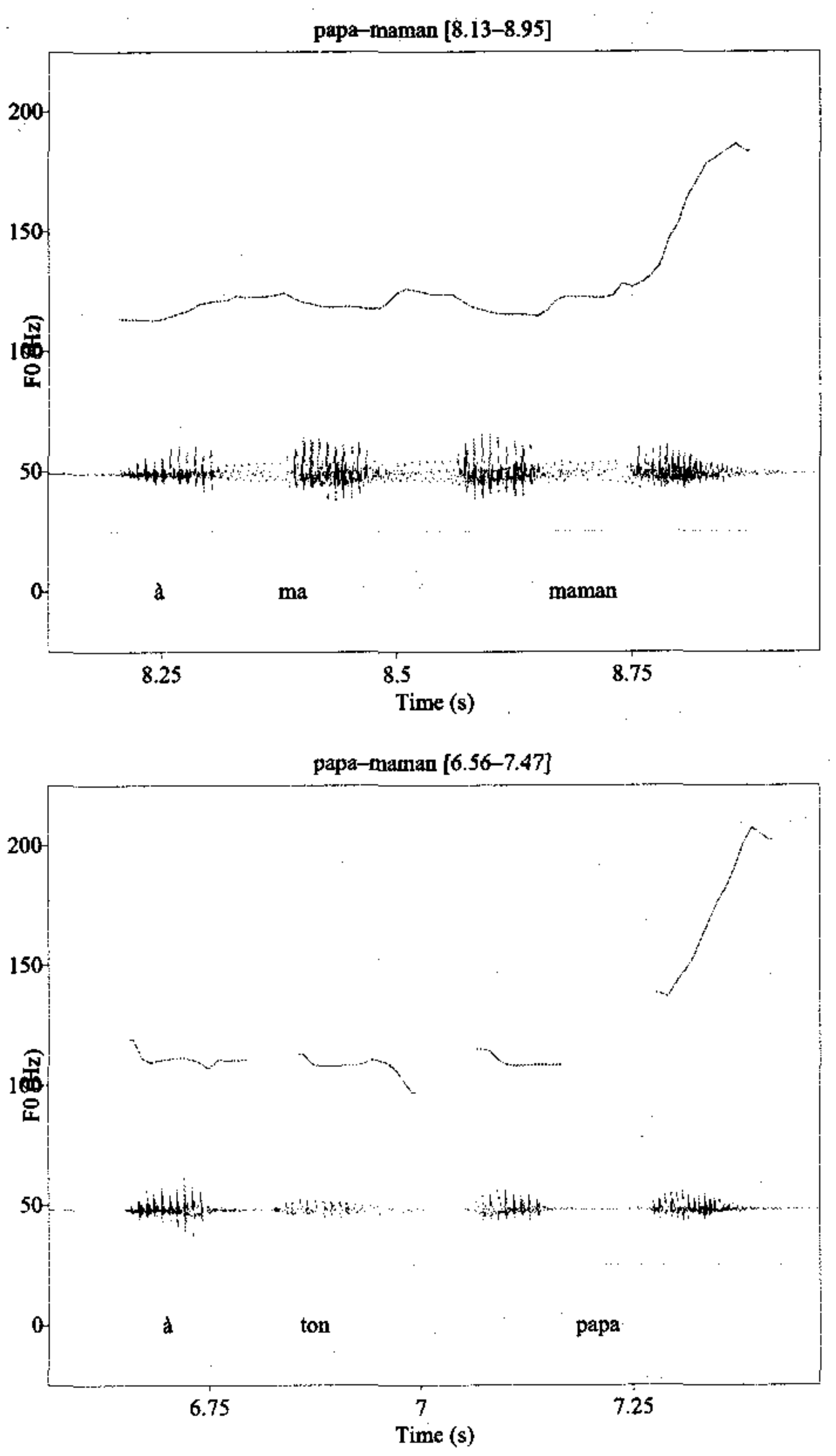

Figure 2. The same utterances as in Figure 1 pronounced as questions.

On the other hand, the utterances "A ma maman!" and "A ma maman?" could be modelled with the same micromelodic profile but with different target point, while "A ton papa!" and "A ton papa?" would also have the same micromelodic profile but which would be different from those of the first pair.

The Momel algorithm derives what I refer to as a *phonetic representation* of an intonation pattern which is neutral with respect to speech production and speech perception since while not explicitly derived from a model of either production or perception it contains sufficient information to allow it to be used as input to models of either process. The relatively theory-neutral nature of the algorithm has allowed it to be used as a first step in deriving representations such as those of the Fujisaki model [19], ToBI [20], [18] or INTSINT [16] .

2.2 INTSINT - a surface phonological representation of intonation

INTSINT is an acronym for **IN**ternational **T**ranscription **S**ystem for **INT**onation.

The first version of this system was proposed in [7] as a prosodic equivalent of the International Phonetic Alphabet and the INTSINT alphabet was subsequently used as a transcription system for intonation patterns in [15] in just over half of the chapters.

INTSINT codes the intonation of an utterance by means of an alphabet of 8 discrete symbols constituting a *surface phonological* representation of the intonation:

T (Top), **M** (mid), , **B** (Bottom),
H (Higher), , **L** (Lower), **S** (Same),
U (Upstepped), **D** (Downstepped).

These tonal symbols are considered *phonological* in that they represent discrete categories and *surface* since each tonal symbol corresponds to a directly observable property of the speech signal.

The tones can be aligned with phonological constituents by means of the following alignment diacritics following the tonal symbol:

[(initial), < (early), : (medial), > (late),] (final)

The relevant phonological constituent with which the tonal segments are aligned can be taken as the sequence of symbols between the following pair of slashes /.../.

The following is an example of a transcription using the International Phonetic Alphabet of a possible reading of the sentence "It's time to go :

M:/jts/T:/tajmtə/D<B]/aə˥/

This corresponds to a *Mid* tone aligned with the middle of the syllable "It's" then a *Top* tone aligned with the middle of the unit "time to" and then a *Downstepped* tone aligned early in the syllable "go" and a *Bottom* tone aligned with the end of the same syllable.

The phonetic interpretation of the INTSINT tonal segments can be carried out using two speaker dependent (or even utterance dependent) parameters of the melody domain.

- **key**: like a musical key, this establishes an absolute point of reference defined by a fundamental frequency value (in Hertz).
- **range**: this determines the interval between the highest and lowest pitches of the utterance.

In the current algorithm [10][11] the tonal segments can be converted to target points, like those generated by the Momel algorithm, using the following equivalences. P(i) in the following formulae refers to the current Pitch target, P(i-1) to the preceding pitch target. Here I assume pitch targets are calculated on a logarithmic scale.

The targets **T**, **M** and **B** are defined 'absolutely' without regard to the preceding targets

T: P(i) := key + range/2
M: P(i) := key
B: P(i) := key - range/2

Other targets are defined with respect to the preceding target:

H: P(i) := (P(i-1) + T) / 2
U: P(i) := (3*P(i-1) + T) / 4
S: P(i) := P(i-1)
D: P(i) := (3*P(i-1) + B) / 4
L: P(i) := (P(i-1) + B) / 2

A sequence of tonal targets such as:

[M T L H L H D B]

assuming values for a female speaker of *key* as 240 Hz and *range* as 1 octave, this would be converted to the following F0 targets:

[240 340 240 286 220 273 242 170]

with appropriate time values derived from the representation of the alignment (omitted here). This sequence of target points can then be used to generate a quadratic spline function modelling the macroprosodic curve of the utterance.

An interesting consequence of this model is that it automatically introduces an asymptotic lowering of sequences such as **H L H**... such as has often been described both for languages with lexical tone and for languages where tone is only introduced by the intonation system, without the need to introduce a specific downdrift or declination component.

The particular values used for calculating the value of **D** and **U** are chosen so that in a sequence **[T D]** for example, the **D** tone is lowered by about the same amount with respect to the **T** as the **H** tone in the sequence **[T L H]**. In many phonological accounts, Downstepped tones are analysed as a High tone which is lowered by the presence of a "floating" low tone, so that the surface tone **[D]** can be considered as underlyingly **[L H]**.

There is now a mailing list devoted to users of Momel and Intsint where it is possible to post questions and comments about the use of these algorithms. The list will also provide updated information about the latest versions and where to find them. This list may be found at the following address:

http://tech.groups.yahoo.com/group/momel-intsint

2.3 Underlying phonological representation of intonation.

As mentioned above, the further we abstract away from the observable acoustic signal the more theory dependent, and hence the more controversial, our representation will become. This is, of course, not specific to speech prosody - there is very little consensus among linguists in general as to the underlying phonological form of words in most languages or as to the underlying syntactic representation of sentences. For this reason, I shall not go into any detail about what such a representation might look like here but the interested reader may consult [8] for one model of an underlying phonological representation of part of the intonation system of English.

Each level I have described is required to be interpretable at adjacent levels - thus a constraint on phonetic representations is that they should be interpretable both acoustically and (surface-) phonologically. It is to be hoped that the type of (semi-)automatic derivation described here, when applied to larger and more varied corpora, may help to make progress in understanding the nature of underlying representations and their relation to prosodic functions.

It may quite possibly never be feasible to establish an exhaustive inventory of prosodic functions for any given language. But with the tools described in this presentation it would be possible for an investigator who has identified acoustic data corresponding to a variety of specific prosodic functions to make an objective assessment of whether a given prosodic model is appropriate for the formal analysis and implementation of these functions.

3 Modeling speech melody.

We assume in this presentation that the fundamental frequency curve is modeled using the Momel and INTSINT algorithms that we have described briefly above and in detail elsewere. [17][8][11] There is, however, nothing in the conception of the ProZed environment which specifically requires this, and other models of fundamental frequency and symbolic coding could equally well be implemented within the same framework.

The first stage of our model is to factor the raw fundamental frequency curve into two components, a *macroprosodic* component corresponding to the melodic pattern of the utterance and a *microprosodic* component corresponding to deviations from this pattern caused by local segmental perturbations. The basic idea behind this is that in speech we can combine different texts with different tunes or intonation patterns. Ideally, our model should separate the components so that when two different texts are pronounced with the same intonation pattern, the macroprosodic component of the two utterances will be the same and when the same text is pronounced with two different tunes, the microprosodic pattern of the two utterances will be the same.

Table 2 shows a sample of the quadratic spline modeling of the raw fundamental frequency curve. The curve is defined by a sequence of target points, each of which is a couple <time (s.), F0 (Hz)>:.

[<0.171, 119>, <0.347, 164>, <0.514, 186>, <0.771, 113>, <1.059, 132>, <1.286, 146>, <1.690, 82>]

Table 2. Sequence of target points <secs., Hz> output by the Momel algorithm from the fundamental frequency curve of the first sentence of the Euroml passage "I have a problem with my water-softener.

Figure 2 shows the raw fundamental frequency curve and the modeled curve defined by the target points in Table 2.

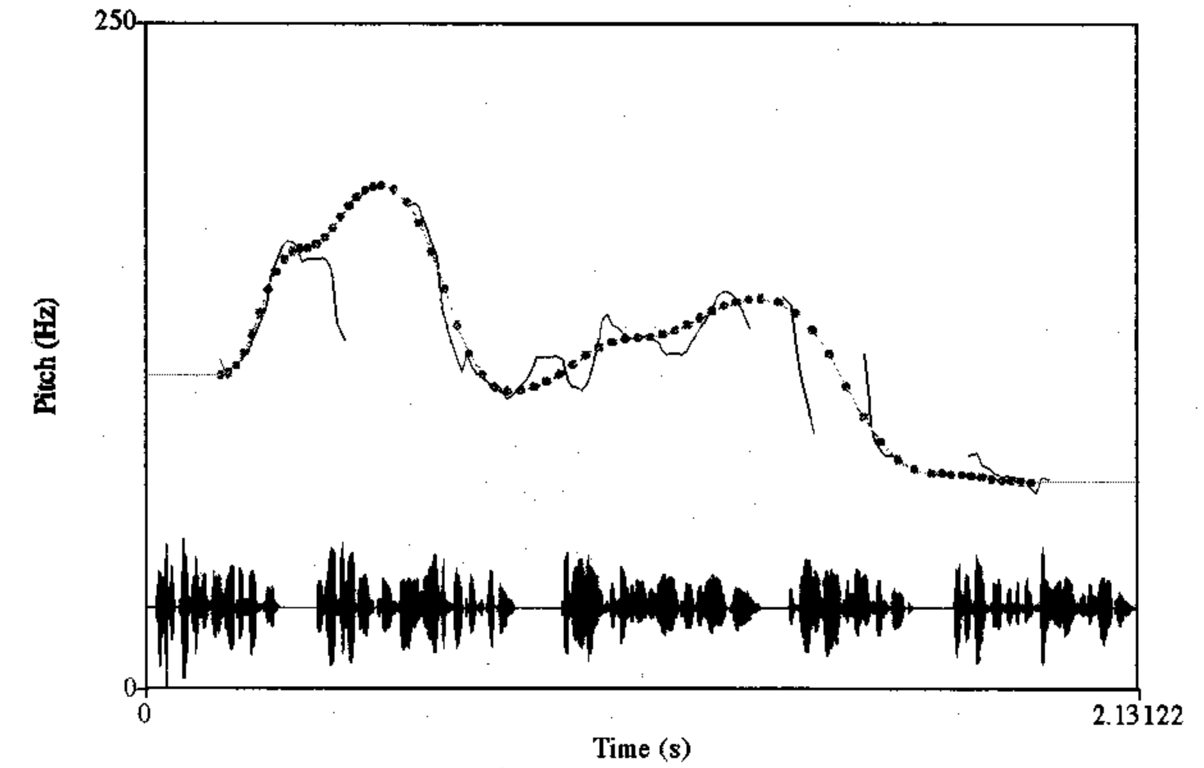

Figure 2. Raw (continuous line) and modeled (dotted line) fundamental frequency of the first sentence of the Euroml passage "I have a problem with my water-softener" defined by the target points in Table 2.

The smooth continuous quadratic spline function shown in Figure 2 provides what I have called a *phonetic representation* of the fundamental frequency curve. I take phonetic representations to be neutral with respect to both speech production and speech perception, unlike models which specifically set out to model either production (e.g. [6]) or perception (e.g. [1]), although it is intended that this representation should capture salient features of both production and perception. Abstracting away from this phonetic representation, we next implement a level of *surface phonological representation* where the melody is represented by a sequence of discrete symbols, using the INTSINT alphabet described above, which codes an intonation pattern using the tonal symbols: **T** (Top), **M** (Mid), **B** (Bottom), **H** (Higher), **S** (Same), **L** (Lower), **U** (Upstepped) and **D** (Downstepped).

For any given pair of parameter values *key* and *range*, there exists an optimal coding of a sequence of target points in terms of mean square error. The first target point cannot refer to a previous target and so it can only be coded as **T**, **M** or **B**, whichever is closest to the observed value, given the current parameters.

The remaining tonal symbols are then selected to give the best fit to the observed data. In the current implementation of ProZed, the complete parameter space for *key* defined by [mean-20Hz...mean+20Hz; step = 1Hz] and for range defined by [0.5 octaves...2.5 octaves; step = 0.1 octaves] is explored without any attempt to optimise the search and the optimal values both for the global parameters *key* and *range* and for the sequence of tonal symbols are selected. et points in Table 2, for example, are coded as the sequence [M T S L H U B] with the parameters *key* = 114 Hz and *range* = 1.102 octaves. Figure 3 shows the comparison of the target points for the whole passage as measured by the Momel algorithm and the same target points coded using the Intsint alphabet and then converted back to numerical values.

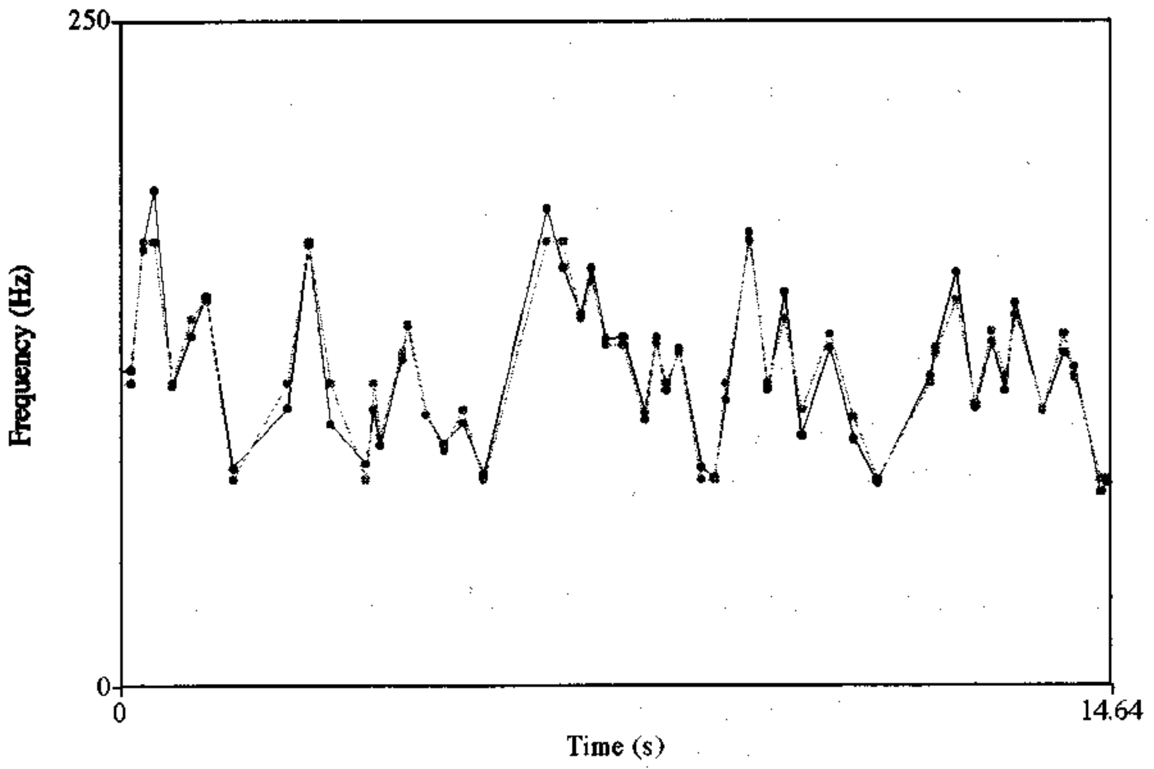

Figure 3: Target points for the complete passage illustrated in Table 1, measured with the Momel algorithm (black) and the same target points coded with the Intsint alphabet then converted back to numerical values (grey).

In [9], it was proposed that the alignment of tonal symbols be defined with respect to the boundaries of the nearest phone using the discrete categories *initial*, *early*, *middle*, *late* and *final*, typically interpreted as representing 0, 25, 50, 75 and 100 percent, respectively, of the phone interval. In our current implementation, we define the timing of the pitch point with respect not to the segment but to a specific *Tonal Unit*. In this illustration, we take the *Tonal Unit* to be the sequence beginning either after an intonation boundary or at the onset of a stressed syllable and continuing up until the next intonation boundary or onset of a stressed syllable. This unit is familiar from numerous descriptions of English intonation, where it is usually referred to as the *stress foot* or the *interstress interval*. Once again, there is nothing in the ProZed environment which requires this particular unit to be defined as the domain for tonal alignment, so that the question of the optimal unit for tonal alignment remains open for empirical investigation. Unlike in [9] where only one tonal symbol per segment was possible, in our current implementation, up to 5 tonal symbols can be defined for each Tonal Unit; the alignment of the tonal symbol with respect to the boundaries of the Tonal Unit are specified, as in the earlier proposal, as *initial*([), *early* (<), *medial* (:), *late* (>) or *final* (]) with the same numerical interpretation as above.

With this coding scheme, the passage from Eurom1 illustrated above can be annotated as in Table 4, where {} indicates boundaries of Tonal Units (including pauses):

```
{}{M}{T>}{S<L:H]}{U<B>}{}{M<}{T<L>}{}{B[H<L:}
{H:U]}{L<L:U>B]}{}{T>}{S<D>}{H<D>}{S]}{L>}{H<
D>}{U[B:S>H>}{T:}{L<}{H<}{L<}{H<L>}{B>}{}{M<
U>}{H<L:H>}{D[H:}{L<H>}{D<B>}
```

Table 4: Intsint coding of the Eurom1 passage given in Table 2. {} indicates boundaries of Tonal Units, the alignment of each tonal symbol is specified as being initial ([), early (<), medial (:), late (>) or final (]) with respect to the boundaries of the Tonal Unit.

4 Combining represenations of melody and rhythm

In [12] [12], I presented a model of English speech rhythm which makes it possible to give an arbitrarily accurate representation of the rhythm of an utterance with a small number of parameters.

It is possible to put these representations or speech rhythm and speech melody together into a single prosodic annotation. Table 5 shows this, using the SAMPA Ascii phonetic alphabet [4]. The representation given in Table 5 can be converted to a low-level phonetic representation, specifying duration and pitch for each segment, and this can then be output for evaluation, to a diphone synthesiser such as Mbrola for example . This makes it possible to use ProZed to test different prosodic models, generated either automatically, as in the procedure described in this text, or by generating or modifying representations like that given in Table 5.

```
<parameter tempo=0.761><parameter quant=50>
<parameter key=114><parameter range=1.102>

{}_{M}aI{T>}h{v@{S<L:H]}prQblm=[1]wIDmaI{U<B>}
wO:t@[3]sQfn@[7]{}_{M<}D@[1]{T<L>}wO:t@[3]levl=
[1]iz[1]{}tu:[4]{B[H<L:}haI[5]{H:U]}n=Di[2]{L<L:U>B]}
@Uv@[1]fl@U[2]ki:ps[2]drIpIN[4]{}_{T>}kUdju:[1]@{S
<D>}reIndZ[3]t@{H<D>}send[2]@n{S]}endZI{L>}nIer[2]
Qn{H<D>}tju:zdeI{U[B:S>H>}mO:nIN[2]pli:z[6]{T:}_{L
<}itsDi:[2]{H<}@Unli:{L<}deI[1]aI[1]k@n{H<L>}m{nId
Z[1]Dis[1]{B>}wi:k[3]{}_{M<U>}aIdbi:{H<L:H>}greItfU
lIfju:kUdk@n{D[H:}f3:m[2]Di:@{L<H>}reIndZmn=t[0]In[
1]{D<B>}raItING[6]{}_
```

Table 5: Combined prosodic annotation for rhythm and melody of the Eurom1 passage given in Table 2 The parameters for rhythm are described in [12] and [12].

The ProZed environment is being implemented as a set of scripts integrated as a plugin for the Praat speech analysis system [2] and will be freely available for research from the author.

CONCLUSIONS

In this paper I have tried to show that the (semi-)automatic modeling techniques we have developed in Aix en Provence are fairly simple to implement and use (in particular from within the Praat environment). It is my hope that the availability of tools such as this will make it possible to carry out large scale analyses of corpora which it would be virtually impossible to undertake manually and will thus contribute to the furthering of our understanding of the nature of prosodic systems.

The possibility of automatically extracting a symbolic representation of both rhythm and melody from the acoustic data opens a number of interesting perspectives for future research. In particular it should make it possible to pursue the evaluation of specific models of prosody on a multilanguage basis and to address objectively the question of the relation between prosodic forms and prosodic functions.

Acknowledgements

Much of the work described here was carried out in collaboration with my two former doctoral students, Cyril Auran and Carol Bouzon (see references).

The opportunity to present this work at this conference was provided by support from the French-Korean bilateral research program STAR as well as from the organisers of the Autumn 2006 conference of the Korean Society of Speech Sciences and Speech Technology.

References

[1] Alessandro C. (d'), Mertens P., 1995. Automatic pitch contour stylization using a model of tonal perception, *Computer Speech and Language* 9(3), 257-288.

[2] Auran, Cyril 2004. Momel-INTSINT [Praat script] downloadable from http://www.univ-lille3.fr/silex/equipe/auran/english/index.html

[3] Boersma, Paul & Weenink, David 2006. Praat: doing phonetics by computer (Version 4.4.23) [Computer program]. Downloadable from http://www.praat.org/

[4] Chan, D.; Fourcin, A.; Gibbon, D.; Granstrom, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; Lindberg, B.; Moreno, A.; Mouropoulos, Senia, F.; Trancoso, I.; Veld. C. & Zeiliger, J. 1995. EUROM- A

Spoken Language Resource for the EU, in *Proceedings of Eurospeech'95*. (Madrid, Spain, September, 1995). (1), 867-870

[5] Dutoit, T, & Leich, H. MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database, *Speech Communication*, Elsevier Publisher, November 1993, vol. 13, n°3-4.

[6] Fujisaki, H., 1998. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. in Fujimura, O. (Ed.). Vocal Physiology: Voice Production, Mechanisms and Functions. Raven Press Ltd., New York, 347-355.

[7] Hirst, D.J. 1987. *La description linguistique des systèmes prosodiques. Une approche cognitive.* Thèse de Doctorat d'Etat, Université de Provence

[8] Hirst, D.J. 1998. Intonation in British English. in Hirst & Di Cristo eds 1998.

[9] Hirst, D.J. 1999. The symbolic coding of duration and alignment. An extension to the INTSINT system. *Proceedings Eurospeech '99.* Budapest, September 1999

[10] Hirst, D.J. 2004. Lexical and Non-lexical Tone and Prosodic Typology. in *Proceedings of International Symposium on Tonal Aspects of Languages*. Beijing, March 2004, 81-88

[11] Hirst, D.J. 2005. Form and function in the representation of speech prosody. in K.Hirose, D.J.Hirst & Y.Sagisaka (eds) *Quantitative prosody modeling for natural speech description and generation* (=*Speech Communication* 46 (3-4)), 334-347

[12] Hirst, D.J. 2006. Analysis by synthesis of speech rhythm: from data to models. in Proceedings 10th KASS Conference, Busan November 2006, 15-20.

[13] Hirst, D.J. & Auran, C. 2005. Analysis by synthesis of speech prosody: the ProZed environment. in proceedings of *Interspeech* 2005.

[14] Hirst, D.J. & Bouzon, C. 2005. The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). in proceedings of *Interspeech* 2005.

[15] Hirst, D.J. & Di Cristo, A. (eds) 1998. *Intonation Systems. A survey of Twenty Languages*. (Cambridge, Cambridge University Press). [ISBN 0 521 39513 S (Hardback); 0 521 39550 X (Paperback)].

[16] Hirst, Daniel, Albert Di Cristo & Robert Espesser 2000. Levels of representation and levels of analysis for intonation. in M. Horne (ed) *Prosody : Theory and Experiment*. Kluwer Academic Publishers, Dordrecht. 51-87

[17] Hirst, Daniel & Robert Espesser 1993. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix 15, 71-85.

[18] Maghbouleh, A., 1998. ToBI accent type recognition. In: *Proceedings ICSLP 98*.

[19] Mixdorff, H., 1999. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings ICASSP 1999*.

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