Analysis by synthesis of speech prosody: the ProZed environment.

Daniel Hirst & Cyril Auran

CNRS, Laboratoire Parole et Langage (UMR 6057),
Université de Provence, Aix en Provence
daniel.hirst@lpl.univ-aix.fr, cyril.auran@lpl.univ-aix.fr

Abstract

This paper presents ProZed, an environment for the multilingual analysis by synthesis of speech prosody. The analysis is based on the symbolic representation of prosodic form without reference to prosodic function. The parameters of the model are at present limited to fundamental frequency and duration but the same framework could be extended to accomodate other parameters such as spectral tilt or voice quality. Each parameter is defined with respect to specific domains and units for long-term and short-term phonetic interpretation with an abstract annotation system corresponding to a level of surface phonological representation. The environment is integrated with the Praat program for analysis and the Mbrola program for synthesis.

1. Introduction.

In recent work [12], [10], 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 this project 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 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 and melody is described 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 distinctive 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, speech 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. Modelling speech rhythm.

The rhythm of speech can be modeled as the interaction of a number of components. In languages with lexical quantity (like Finnish) there is a lexical specification of length. In other languages, segmental duration is influenced by the accentual structure of the utterance and by longer-term more subtle variations of global duration called *tempo*.

We propose here, as an illustration of our methodology, a model of rhythm for British English that we have implemented with the ProZed environment.

In [8], as a first approximation, a simple scalar feature of length was proposed, with 5 degrees*: extra-short, short, normal, long* and *extra-long*. This feature was assumed to apply directly at the level of the segment.

There is, however, evidence that lengthening can be better handled at a higher level. Eriksson [6] showed that, for a number of European languages, when the stress foot is taken as the domain of lengthening, its duration can be satisfactorily modeled as a linear function of the number of constituent syllables, with a language specific offset of approximately 100 ms for syllable-timed languages and of approximately 200 ms for stress-timed languages. [1] and [11] argue that the domain of lengthening in (British) English is not the stress-foot but, following [13], the *narrow rhythm unit,* defined as a prosodic unit beginning with a stressed syllable and ending at the next word boundary. Any unstressed syllables not part of a narrow rhythm unit are classified as belonging to the *anacrusis*, which is not lengthened.

As in [8], lengthening in this model is taken to be a simple scalar feature. In this implementation, however, the lengthening is taken to apply at the level of a specific domain that we call the *rhythm unit*. Rhythm units are delimited by word boundaries and by the onset of the (primary or secondary) lexically stressed syllable. They thus correspond to, both narrow rhythm units and anacruses. Monosyllabic

words in this model, will consequently all constitute single rhythm units; polysyllabic words will consist of one or more rhythm units depending on the position and number of stresses. Thus a word like "communication" /kəˌmjuːnɪˈkeɪʃn̩/ will be taken to consist of three rhythm units /kə/, /mjuːnɪ/ and /keɪʃn̩/. Another difference with [8] is that whereas in that presentation, the scalar values corresponded to five discrete categories: *extra-short, short, normal, long* and *extra-long*, in this model there are only two discrete categories: *unlengthened* and *lengthened* where *lengthened* is associated with a scalar factor *k*. This model allows us to code the observed duration patterns of an utterance as a combination of two factors: a global factor of *tempo* and a local scalar degree of lengthening *k* applying to each rhythm unit. The lengthening factor for a given rhythm unit is calculated by comparing with the sum of the trimmed mean values of the constituent phonemes corrected by a global value for *tempo*, where the trimmed mean corresponds to the mean of the central 80% values of the quantile range.

The coding is carried out iteratively. The lengthening factor *k* for each rhythm unit is initially set to 0 and the tempo *t* is set to the sum of observed phone durations divided by the sum of predicted durations. The predicted duration $\hat{d}_{ru}$ of a rhythm unit containing *m* phonemes is then adjusted to:

$$\hat{d}_{ru} = \left(\sum_{i=1}^{m} \left(\bar{d}_{i/p} \right) + k * q \right) * t \qquad (1)$$

where $\bar{d}_{i/p}$ corresponds to the mean duration of each constituent phoneme *p*, *k* is the scalar lengthening factor, *q* is a quantal duration unit which we set to the (trimmed) overall mean duration of all phones and *t* is the current value of tempo. At each iteration, the lengthening coefficient *k* of the rhythm unit with the largest positive error of prediction [d_{ru} - $\hat{d}_{ru}$] is incremented and the value of the tempo coefficient *t* is recalculated. This is reiterated until there are no positive errors greater than a given proportion (e.g. 0.75) of the quantal factor *q* multiplied by the current tempo *t*.

As an illustration of this modelling technique, a recording from the Eurom1 corpus [4] was analysed. In Table 1, the rhythm units are transcribed orthographically. When a word contains more than one rhythm unit as in "arrange" or "engineer", these are separated by hyphens. The figure in square brackets after each rhythm unit corresponds to its estimated lengthening factor *k*, the parameters *t (tempo)* and *q (quant);l* are specified at the beginning of the passage.

<parameter tempo=0.761><parameter quant=50>
I have a problem[1] with my water[3] softener[7]. The[1] water[3]-level[1] is[1] too[4] high[5] and the[2] over[1]-flow[2] keeps[2] dripping[4]. Could you[1] a-rrange[3] to send[2] an engi-neer[2] on Tuesday morning[2] please[6]. It's the[2] only day[1] I[1] can manage[1] this[1] week[3]. I'd be grateful if you could con-firm[2] the a-rrangement in[1] writing[6].

Table 1. A sample passage from the Eurom1 corpus coded for duration using the automatic coding scheme described in the text. Hyphens and spaces separate *rhythm units*, numbers in square brackets correspond to the scalar lengthening factor *k* as applied to the preceding rhythm unit.

In this example it is assumed that the whole passage is the domain for the parameters *t* and *q*. It is, however, perfectly possible to split a passage into several separate domains and to optimise the coding for each sub-passage with specific values of *t* and *q* for each domain.

Figure 1 shows the comparison of observed and fitted durations for the passage coded in Table 1. As can be seen, the fit is very close (r = 0.9946) and could be adjusted to an arbitrary degree of granularity by modifying the threshold conditions on the iterative process.

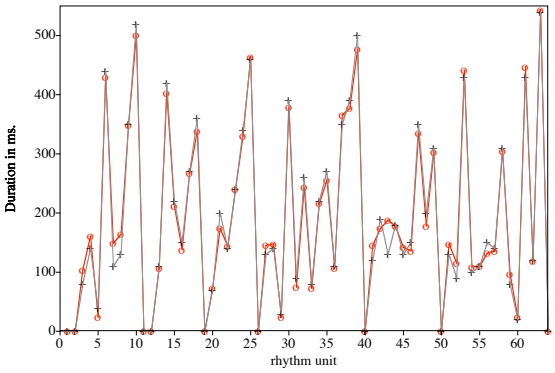

Figure 1. Observed (+) and fitted (o) durations of rhythm units for the Eurom1 passage in table (1) above.

3. Modelling speech melody.

We assume in this presentation that the fundamental frequency curve is modeled using the Momel and INTSINT algorithms that we have described in detail elsewhere [12], [9], [10]. There is, however, nothing in 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 modelling 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)>:.

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

[<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 Eurom1 passage "I have a problem with my water-softener."

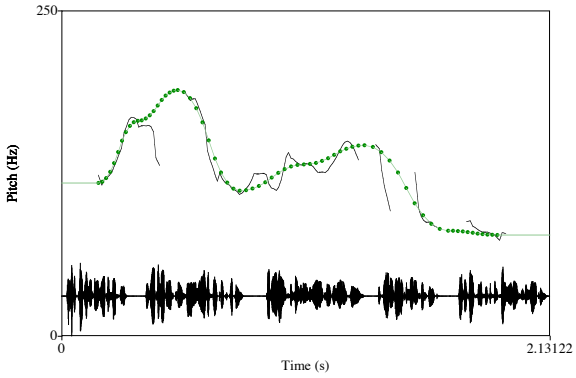

Figure 2. Raw (continuous line) and modeled (dotted line) fundamental frequency of the first sentence of the Eurom1 passage "I have a problem with my water-softener" defined by the target points in Table 2.

The smooth continuous quadratic spline function shown in Figure 2 provides what we have called [12] a *phonetic representation* of the fundamental frequency curve. We 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. [7]) or perception (e.g. [1]), although it is intended that this representation should capture salient features of both production and perception (for an application of the Momel algorithm to the Fujisaki model of pitch production cf. [14]).

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 [12] 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).

The phonetic interpretation of these tonal symbols is established with reference to two parameters: *key* and *range*, which, like the parameter *tempo* for rhythm are defined within a global domain. Given the global values of these two parameters, each tonal symbol can be interpreted as defining a unique target value, either absolutely (for the symbols **T**, **M** and **B**) or relative to the preceding target point for the other symbols according to the formulae in Table 3, where $P_{i/ts}$ represents the value of target point *i* with a tonal symbol *ts*. The basic idea behind this interpretation is that the absolute tonal symbols **T**, **M**, **B** refer to the top, middle and bottom of the current pitch range, respectively, whereas the relative symbols define a pitch interval which is a given proportion of the distance between the preceding target point and the top or bottom of the pitch range.In our current implementation these formulae are interpreted on a logarithmic scale.

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.

$$P_{i/T} = key + range/2$$
$$P_{i/M} = key$$
$$P_{i/B} = key - range/2$$
$$P_{i/H} = P_{i-1} + a_H(P_T - P_{i-1})$$
$$P_{i/S} = P_{i-1} + a_S(P_B - P_{i-1})$$
$$P_{i/L} = P_{i-1} + a_L(P_B - P_{i-1})$$
$$P_{i/U} = P_{i-1} + a_U(P_T - P_{i-1})$$
$$P_{i/D} = P_{i-1} + a_D(P_B - P_{i-1})$$

a_H and a_L are typically set to 0.5, a_U and a_D to 0.25 and a_S to 0.

Table 3. Formulae for the phonetic interpretation of INTSINT symbols.

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 is selected. The target 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.

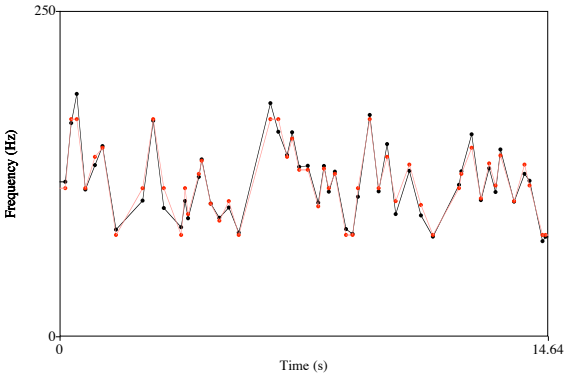

Figure 3: Target points for the complete passage illustrated in Table 1, measured with the Momel algorithm (black) and the same target points coded with the Intsint alphabet then converted back to numerical values (grey).

In [8], 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 [8], 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 [8], 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.

We are now in a position 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 [15].

```
<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>}w
O:t@[3]sQfn@[7]{}_{M<}D@[1]{T<L>}wO:t@[3]levl=[1]i
z[1]{}tu:[4]{B[H<L:}haI[5]{H:U]}n=Di[2]{L<L:U>B]}@U
v@[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{nIdZ[1]Dis[1
]{B>}wi:k[3]{}_{M<U>}aIdbi:{H<L:H>}greItfUlIfju:kUdk
@n{D[H:}f3:m[2]Di:@{L<H>}reIndZmn=t[0]In[1]{D<B>}r
aItING[6]{}_
```

Table 5: Combined prosodic annotation for rhythm and melody of the Eurom1 passage given in Table 2.

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. 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.

The ProZed environment is implemented as a set of scripts integrated with the Praat speech analysis [2] and is freely available for research from the authors.

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] Boersma, P. & Weenink, D. 2005. Praat. Doing phonetics by computer. [computer program]. Version 4.3.04 Retrieved March 31, 2005 from http://www.praat.org/

[3] Bouzon, C. & Hirst, D.J. 2004. Isochrony and prosodic structure in British English. in Bel, B. & Marlien, I. (eds): *Proceedings of the Second International Conference on Speech Prosody*, Nara, Japan, 223-226.

[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] Eriksson, A., 1991. *Aspects of Swedish Speech Rhythm. Doctoral dissertation,* University of Göteborg: Sweden

[7] 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.

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

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

[10] 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*. (= special issue of *Speech Communication*).

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

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

[13] Jassem, W. 1952. *Intonation in Conversational English*. Warsaw, Polish Academy of Science.

[14] Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP2000)*. Vol. 3. Istanbul, 1281-1284.

[15] Wells, J.C., 1997. 'SAMPA computer readable phonetic alphabet'. In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B.
http://www.phon.ucl.ac.uk/home/sampa