CODING THE F0 OF A CONTINUOUS TEXT IN FRENCH : AN EXPERIMENTAL APPROACH.

Daniel Hirst, Pascale Nicolas and Robert Espesser

Institut de Phonétique d'Aix,
URA CNRS 261 Parole et langage.
Aix en Provence, France

ABSTRACT

An algorithm for the automatic modelling of fundamental frequency curves as a quadratic spline function was applied to a continuous text in French. The output of the model, a sequence of target pitches <ms; Hz>, was then coded using four successively more complex models which were subsequently used to generate synthetic versions of the recording, each version respecting the statistical distribution of the modelled target pitches. The resulting recordings were evaluated subjectively by native speakers. The two more complex codings obtained over 80% of the score obtained by the resynthesis of the text using the measured targets.

1 INTRODUCTION

A number of speech synthesis systems are available today for several different languages, capable of intelligibly synthesising isolated sentences. Results for continuous texts, however, are far less satisfactory. It is generally agreed that one of the principal weaknesses of such systems is the inadequate modelling of prosodic parameters : fundamental frequency, intensity and segmental duration.

In this paper we present the preliminary results of a project investigating the fundamental frequency structure of continuous texts in French.

2 METHOD

The research makes use of an automatic fundamental frequency modelling program MOMEL [6] combined with the PSOLA technique for time domain prosodic modification of speech [2].The F0 modelling program uses a dissymetric version of robust regression to provide an optimal fit for a sequence of parabolas,.factoring the F0 curve into two components, a microprosodic profile and a macroprosodic profile [1]. The output of the program is in the form of a sequence of target-points <ms; Hz>. These target-points can subsequently be used to generate a quadratic spline function [3] which is then directly usable as input for PSOLA resynthesis. For very high quality synthesis the microprosodic profile can be reintroduced, although owing to the high quality of the PSOLA synthesis, even without microprosodic correction, the resynthesis using this technique is practically indistinguishable from an original recording as far as the intonation is concerned.

3 CORPUS

The corpus used in the experiment was recorded from an introduction to science for French children. The text consists of 3 paragraphs, 8 sentences, 140 words and 232 syllables and develops a single topic "the atom".

The text, presented in normal orthography with its original punctuation, was recorded in an anechoic chamber by 6 subjects : 3 male and 3 female, all native speakers of French. The recording used for the experiment described here was that of a 30 year old female.subject whose reading was considered the most satisfactory of the six, presenting a harmonious rhythm and no hesitations. The complete recording including pauses lasted 55secs.

4 ANALYSIS

The fundamental frequency of the recording was analysed by means of spectral comb analysis [8], and the F0 was subsequently modelled by the automatic modelling program described above resulting in a set of 170 target points. Of these 170 values 3 were deleted and 3 others added manually after visual and auditory evaluation of the modelled curve. The resulting values, (the first 40 are illustrated in Figure 1) constitute the reference set **A0**.

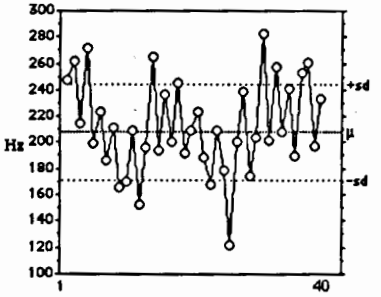

Figure 1 : first 40 target values from the reference set **A0**.

5 PROSODIC CODING

Four approximations to the original prosody of the text were obtained.by successively more complex prosodic coding of the different target points. In this analysis only the F0 values of the target-points were modelled : the time values being taken as given.

5.1. Model A1

The first approximation consisted of a sequence of values assigned randomly but with the same statistical distribution (i.e. the same mean and standard deviation) as the reference set **A0**. :

A1	N	mean F0	s.d.
	170	199	37.1

This approximation was introduced to test the possibility that the only relevant function of fundamental frequency variation is to avoid monotony.

5.2. Model A2

In a second approximation, each value was coded as either Higher (**H**) or Lower (**L**) than the preceding target point. This coding was intended to introduce a distinction between relatively high and low points of an F0 curve corresponding to the distinction used in a number of phonological models of intonation between High and Low tones constituting pitch accents [9],[3],[4].In a preliminary attempt, a set of target points was generated such that the intervals between successive points had the same statistical distribution as those corresponding to similarly coded points of the reference set **A0**. Informal listening showed however that such a model was very unsatisfactory since there was a tendancy for a sequence of values to drift into very high or very low regions, beyond the normal range of the subject's voice. To counteract this, a form of asymptotic declination needs to be introduced [7]. An extremely simple model of declination is given by the formula : $h_i = \sqrt{h_{i-1} * h_A}$ [4] where h_i is a pitch target and h_A is an asymptotic value. This formula was originally intended to model pitch lowering but the same formula can be used for both lowering and raising assuming simply distinct asymptotic values. Estimates of h_A can be then made simply. from each successive pair of values using the formula : $h_A = (h_i)^2 / h_{i-1}$ The successive values can then be generated using the declination formula and an asymptotic value with the same statistical distribution as that of similarly coded values of the reference set **a0**

A2 code	N	mean asymptote	s.d.
L	82	166	57.1
H	87	255	100.3

5.3. Model A3

In a third approximation, the text was segmented manually into Intonation Units on the basis both of textual content and of the F0 contour. The highest point in each Intonation Unit was then coded Top (**T**), the lowest point Bottom (**B**). The first point of each unit was coded Mid (**M**) if not already coded. Other points were first coded Higher (**H**) and Lower (**L**), as in **A2** and then recoded so that an **H** immediately followed by **H** or **T** was recoded as Down (**D**) while **L** immediately followed by **L** or **B** was

recoded as U. This coding corresponds to a transcription using INTSINT (an INternational Transcription System for INTonation) as proposed recently [5] in an attempt to set up a system capable of transcribing significant pitch patterns in any language. The coding incorporates two types of symbols : absolute symbols T, M and B, and relative symbols ; H, D, U and L. Target points coded with absolute symbols were assigned F0 values having the same statistical distribution as similarly coded points of the reference set A0.

A3	code	N	mean F0	s.d.
(absolute)	B	17	146	19.6
	M	13	204	23.7
	T	17	263	19.0

Targets coded with relative symbols were assigned values using the declination formula already used for A2 and an asymptotic value with the same statistical distribution as that of similarly coded values of the reference set A0:

A3	code	N	mean asymptote	s.d.
(relative)	L	44	152	24
	U	10	208	28
	D	24	147	20
	H	45	262	39

5.4. Model A4

The absolute values coded in A3 all show a fairly large standard deviation reflecting considerable variability. The final approximation A4 tested here was designed to limit the variability of T and B by restricting these symbols to target points respectively higher than 281 Hz and lower than 141 Hz.

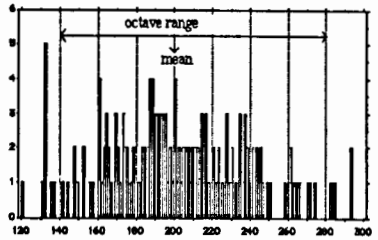

Figure 2 : distribution of target points from the reference set A0.

These thresholds were set to include only values which were beyond an octave range centred around the mean F0 of the recording. As can be seen from the frequency distribution in Figure 2 these values isolate fairly well the extreme values of the distribution which also correspond well to the beginning and end-points of paragraphs.

A4	code	N	mean F0	s.d.
(absolute)	B	10	133	5.5
	M	13	204	23.7
	T	4	288	5.8
	code	N	mean asymptote	s.d.
(relative)	L	51	149	24
	U	10	208	28
	D	24	147	20
	H	58	279	58

6 EVALUATION

In order to reduce the quantity of material to be evaluated, only the first two thirds of the text were used for the evaluation test. Three different series of recordings were obtained, each containing one version generated from the reference set A0 as well the four approximations A1, A2, A3 and A4 described above. No corrections were made for microprosodic effects.The five recordings in each series were presented in random order. The actual set of target values for each of the four models was different for each series but respected the statistical constraints described above. The first series was used as practice material. Subjects were asked to listen to the 5 recordings which, they were told, had undergone a number of different treatments which might affect the way the readings sounded. Subjects were then asked to listen to the second and third series of recordings and to underline portions of the text which they felt sounded least satisfactory. At the end of each recording subjects were asked to attribute a global score out of 20 reflecting their overall satisfaction with the recording.

7 RESULTS

12 subjects took part in the evaluation test, all students or personnel of the Université of Provence. None of the subjects had any particular training in phonetics. An analysis of variance of the subjects' scores showed a significant difference between the scores for the various models ($F(4,110) = 13.286$, $p = 0.0001$) but no significant difference between the two series ($F < 1$) and no significant interaction between model and series ($F < 1$). The reference set A0 received a mean score of 16.5. The four approximations received the following mean scores :

	A1	A2	A3	A4
score	8.5	11.1	13.4	13.6

8. DISCUSSION

Listeners showed a clear preference for the binary (H,L) coded targets (A2) compared to the random distribution (A1). They also showed a clear preference for the two versions of INTSINT tested as compared to the other approximations. The version of INTSINT incorporating a threshold (A4) received a slightly better score than the other version (A3) but the difference was not significant here. Both versions of INTSINT attained more than 80% of the score attributed to the reference set A0. This suggests that while improvement is still possible, these two models provide a very reasonable approximation to the estimate values. Further experimentation will be necessary, however, to decide whether the incorporation of a threshold provides a distinct improvement to the coding.

It is worth emphasising that the only manual intervention in the coding concerned the determination of the position of the boundaries of Intonation Units. Once these boundaries are determined, the coding used in both models A3 and A4 is entirely automatic. Inspection of the means of both the absolute and the relative values for the two versions of INTSINT suggests an interesting further generalisation. The mean values for T, M and B are quite close to the mean asymptote values for H, U and L/D respectively, with L and D seeming to have identical .asymptotic values. An extremely simple approximation to these values can be obtained from the mean F0 and the two extreme values covering an octave range centred on the mean, the same values which were used as threshold values in model A4. This means that the phonetic implementation rules described above could be implemented using a single individual parameter : the mean speaker F0 together with a standard deviation of say 10%. It remains to be seen, however, how far such a model can be generalised to other speakers and other languages.

9 REFERENCES

[1] Di Cristo, A. & Hirst, D.J. (1986) "Modelling French micromelody : : analysis and synthesis." *Phonetica* 43, 11-30

[2] Hamon, Moulines & Charpentier (1989) "A diphone system based on time domain prosodic modifications of speech". *Proc. Int. Conf. Assp.*, 239-241.

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

[4] Hirst, D.J. (1987) *La description linguistique des systèmes prosodiques : une approche cognitive.* Thèse de Doctorat d'Etat, Université de Provence

[5] Hirst, D.J. & Espesser, R. (1991) "Automatic modelling of fundamental frequency." *Travaux de l'Institut de Phonétique* 15

[6] Hirst, D.J. & Di Cristo, A. (in press) "A survey of intonation systems." in Hirst & Di Cristo *Intonation Systems : a Survey of Twenty Languages.* (Cambridge University Press; Cambridge)

[7] Liberman, M & Pierrehumbert, J (1984) "Intonational invariance under changes in pitch range and length." in Aronoff & Oehrle (1984) *Language Sound Structure* 157-253

[8] Martin, P (1983) "Real time fundamental frequency analysis using the spectral comb method." *Proc. Phon. Sci. X*, Volume 2, 284-287.

[9] Pierrehumbert, J. (1980) *The Phonology and Phonetics of English Intonation.* PhD thesis; MIT.