Automatic Modelling Of Fundamental Frequency Using A Quadratic Spline Function.

Article · February 1999

Source: CiteSeer

CITATIONS
221

READS
420

2 authors:

Daniel Hirst
French National Centre for Scientific Research (CNRS) & Aix-Marseille University
174 PUBLICATIONS **3,085** CITATIONS
SEE PROFILE

Robert Espesser
French National Centre for Scientific Research
114 PUBLICATIONS **1,766** CITATIONS
SEE PROFILE

Travaux de l'Institut de Phonétique d'Aix
volume 15, pages 75-85

AUTOMATIC MODELLING OF FUNDAMENTAL FREQUENCY USING A QUADRATIC SPLINE FUNCTION.

Daniel Hirst & Robert Espesser

CNRS (URA 261), Institut de Phonétique d'Aix, Université de Provence.

Introduction.

There have been a number of different implementations of phonological/phonetic models of intonation designed to derive an acoustic output (F_0 curve) from a symbolic input (for recent overviews cf. Hirst 1991, Monaghan 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 article we shall reserve the term 'model' for attempts to solve both the coding problem and the synthesis problem together:

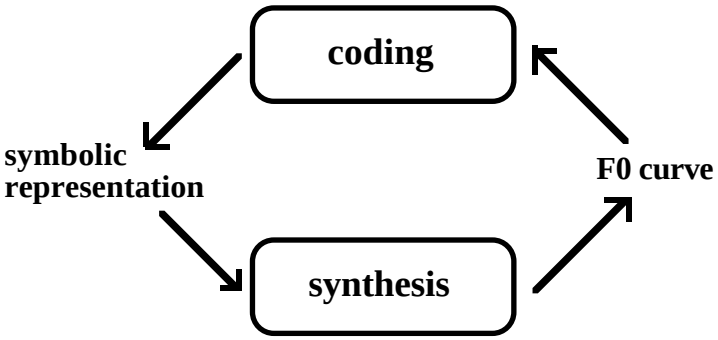

Figure 1 outline of a model of fundamental frequency.

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, the modelling problem has received surprisingly little attention from workers in the field. A few exceptions are Scheffers (1988) who describes a technique for obtaining an automatic piece-wise linear approximation of an F0 curve and Taylor & Isard (1992) who describe a model analysing an F0 curve as a linear sequence of 3 primitive contours: rise element, fall element and connection element.

In this paper, after a discussion of some of the background and assumptions behind our work we present an algorithm which has been developed in the Institut de Phonétique d'Aix and which constitutes, so far as we are aware, the first automatic programme of its kind.

1. Background and assumptions

The algorithm we describe here builds on a number of underlying assumptions about an appropriate form for different possible levels of representation for intonation. There is considerable difference of opinion concerning these levels of representation (for discussion cf. Hirst 1991, 1992). There does, however, appear to be a certain degree of agreement that raw fundamental frequency curves can be analysed as the superposition of two fairly independent components: a microprosodic component caused by the nature of the individual phonematic

segments of the utterance and a macroprosodic component reflecting the choice of intonation pattern for the utterance. One of the first steps in modelling an F0 curve, then, is to attempt to factor out the raw curve into its two components.

- 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."
(ii) "C'est un mot, _____."
(ii) "C'est un mot, _____?"

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. As can be seen in the following figure illustrating the voiced stops for one speaker, the relationship between the F0 on the consonant and on the surrounding vowels appears to be fairly linear.

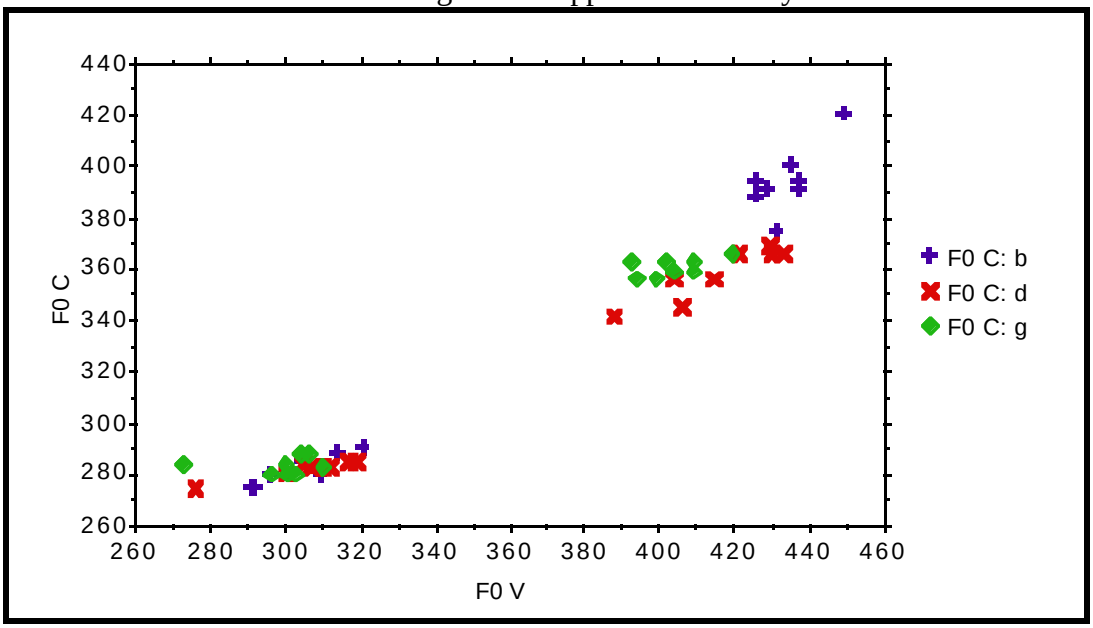

Figure 2: fundamental frequency at the centre of voiced consonants /b/ /d/ /g/ compared to that of surrounding vowels /a/. Data from Nicolas (1989)

In fact, Nicolas showed that a logarithmic regression gave slightly better predictions than a linear one for most speakers and most consonants although the differences between linear and logarithmic predictions were very small. We conclude then that the raw f0 curve can be factored out into a microprosodic component which, at least at a first approximation, can be treated as a multiplicative factor superimposed on a (linear or logarithmic) macroprosodic component.

- 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 (Di Cristo & Hirst 1986).

We have found that the F0 curves for sentences of this type ("Molly may marry Larry", "Oui il a loué l'île aux lilas.") can be very closely modelled by a quadratic spline function and we have argued elsewhere (Hirst 1980, 1983, 1987) that a function of this type allows us to treat a sequence of target points as an appropriate *phonetic* representation for F0 curves.

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 have the advantage that they interpolate monotonically between points of which the first derivative is zero. 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[1], that is that the inflection point is always situated halfway between two adjacent targets.

A recent study ('t Hart 1991) has suggested that our attempt to synthesise F0 curves with parabolas rather than with straight lines (as has been standard practice for several years in Eindhoven) is misguided since subjects cannot hear the difference anyway. This result calls for a few comments.

(i) 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:

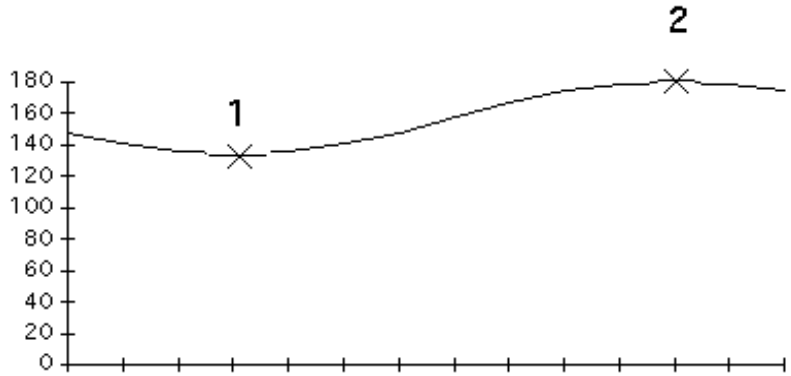

Figure 3: Coding an F0 rise using a quadratic spline function.

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

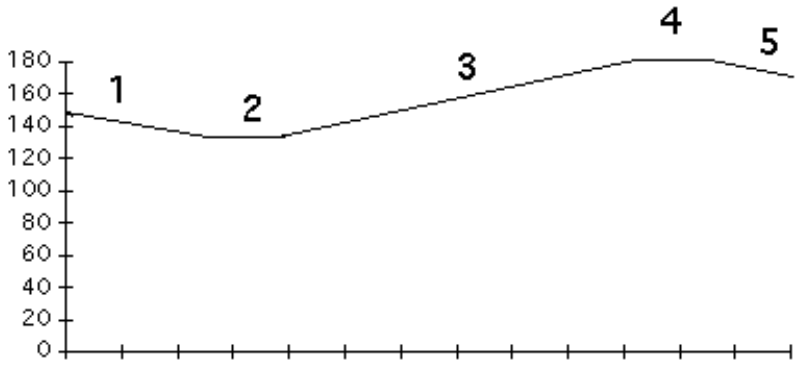

Figure 4: Coding an F0 rise as a sequence of straight lines.

The economy of such a representation is not evident.

[1] 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.

(ii) We do not consider that 't Hart has demonstrated conclusively that straight-line synthesis is always just as good as quadratic spline functions. Since quadratic spline functions give a closer approximation to real F0 curves than do straight lines, it is, in our opinion, quite possible that these differences will be appreciable under certain circumstances. Even though subjects may claim that they are unable to distinguish certain stimuli it is well known that under certain circumstances these stimuli may give rise to different reactions.
(iii) Finally, we might conclude that the very fact that the automatic coding algorithm 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 this output to feed into a straight-line synthesis system but the direct comparison between observed F0 and modelled F0 would no longer be possible.

2. An algorithm for automatic coding of F0 curves

The 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. Since real microprosodic components both raise and lower this curve, a preliminary stage is needed (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) then selects a partition of these target candidates. The final stage (stage 4) reduces the candidates within each partition to a single target.

Modal Regression

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. In 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. Thus while a distribution only has one mean it may have a number of different modes, and even when there is only one mode its value may be dependent on the choice 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. We can define modal regression, then, as selecting the parameters of a given function such that it is 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 tell us how we are to go about selecting the parameters to fulfill this aim. The task will be further complicated by the fact that like the mode of a distribution there may be more than one value for the parameters of a modal regression. One method for selecting these modal parameters is presented in §3.2 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 describe below is designed to eliminate, all other microprosodic effects consist of a lowering of the "underlying" macroprosodic curve which we propose to model as a quadratic spline function. We consequently introduce the further 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) pre-processing 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 §4 below.

The fundamental frequency curves input to the algorithm are detected using a comb function (Martin 1981, Espesser 1982) and are sampled every 10ms with 0 for values in unvoiced zones.

2.1 pre-processing of f0

All values which are 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 pre-processing has essentially the effect of eliminating one or two values (i.e. about 10 to 20ms) at the onset of voicing.

2.2 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. Typical values for the thresholds are 50 Hz and 500 Hz respectively.

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 coefficients

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

where :

$$t = -b/(2c)$$

$$h = a + bt + ct^2$$

If *t* 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.

2.3 partitionning of target candidates

Within a moving window of length *R* (typically 200ms) centred on each instant *x, td(x)* and *hd(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 is then obtained by weighting these distances :

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

where :

$$wd = \frac{1}{mean(dt(x))}$$

and

$$wh = \frac{1}{mean(dh(x))}$$

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

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

2.4 reduction of candidates

Within each segment of the partition, 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.

3.Evaluation

The algorithm described above uses 5 independent parameters :

- minimum value for F0 *[hzmin]*
- maximum value for F0 *[hzmax]*
- 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 microprosodic configurations) of French, spoken by ten subjects (5male, 5 female).

S1 : "La pipe de Jean s'est cassée en tombant de ta gabardine."

S2 : "La fille de Charles Sablon a voulu un petit chien en guise de cadeau."

The following figures illustrate the application of the algorithm to sentence S2 of corpus VNV.

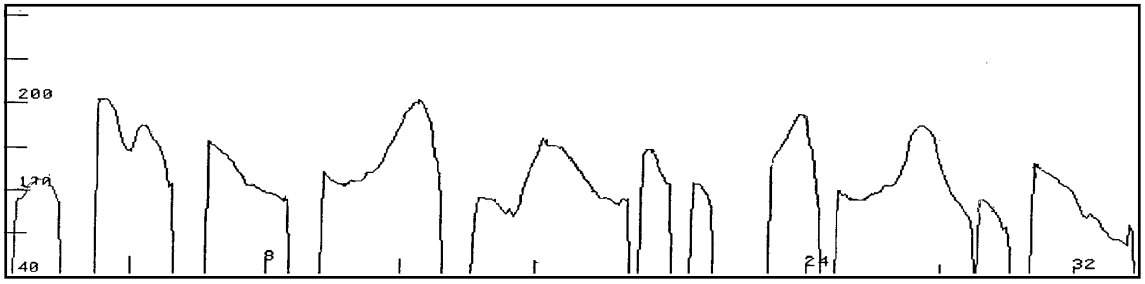

Figure 5. Fundamental frequency for the sentence : "La fille de Charles Sablon a voulu un petit chien en guise de cadeau."

In the following figure 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.

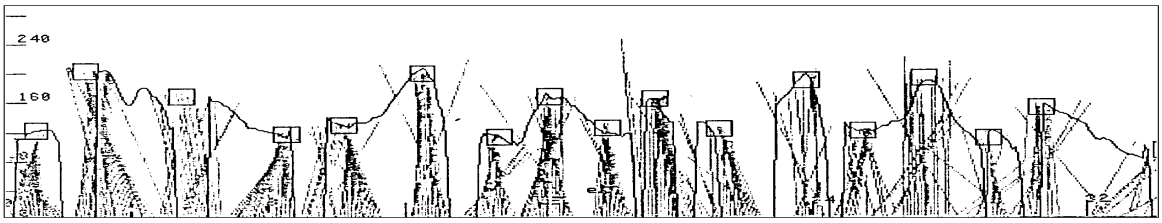

Figure 6. Estimates of targets by asymmetrical modal quadratic regression (see text).

The squares in Figure 6 correspond to the final estimates of the targets after partionning the different candidates. The resulting modelled curve is illustrated in Figure 7 by the continuous grey line.

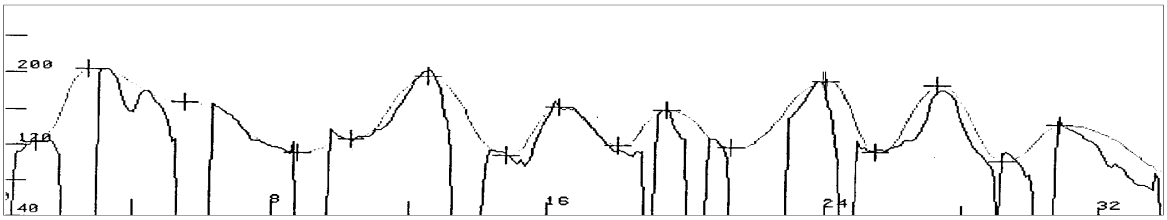

Figure 7. Fundamental frequency curve and quadratic spline model.

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 calculated 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 correspondence between the two types of evaluation was observed.
The minimum and maximum values for F0 were found to be quite robust so that it was possible to fix the same values for all ten speakers:

- *[hzmin]* : 50
- *[hzmax]* : 500

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

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

The algorithm was subsequently applied with its parameters fixed to the above values to two other corpora of rather different nature :
- Corpus *METEO* (Nancy)
Man/machine dialogue concerning the weather conditions in the region of Nancy.
10 speakers. About 2 minutes per speaker.
- Corpus *TAIX* (Aix en Provence)
Three continuous texts.
4 speakers. (2 male, 2 female). About 5 minutes per speaker.
The subjective and objective evaluation techniques described above were applied to the recordings of 2 speakers of the Corpus METEO and 1 text of 1 speaker of the corpus *TAIX*. (Hirst, Nicolas & Espesser 1991).
The total error rate for these recordings (missed targets and false targets combined) although slightly higher than that for the corpus VNV remained quite reasonable (around 5%). The mean distance was also quite close to that observed on the first corpus.

Table 1 : summary of subjective and obective evaluation of the algorithm for 3 corpora.

CORPUS	mean distance	number of targets	number of errors	total length
VNV	0,0622	284	6 (2,1%)	49s
METEO				
Speaker 6	0,0517	429	24 (5,2%)	119s
Speaker 8	0,0454	343	11 (3,2%)	93s
TAIX				
Text 1 Sp. 1	0,0619	161	8 (5,0%)	55s

Conclusion

While the algorithm described above is still somewhat less than perfect, it does in our opinion constitute an interesting first approximation to a working 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 (Najim & Hirst this volume) and is apparently fairly robust.

References

Bailly, G., Benoît, C. & Sawallis, T. (eds) (1992).*Talking Machines : Theories, Models and Designs*. Elsevier Science.

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

Espesser, R. (1982) "Un système de détection du voisement et de F0." *TIPA* 8, 241-261

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

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

Hart, J. ('t) (1991) “F0 stylisation in speech : straight lines versus parabolas." *J. Acoust. Soc. Am.* 6, 3368-3370.

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 Bailly Benoît & Sawallis (eds) (1992), .

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.

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.

Monaghan, A. (1992) "Generating synthetic prosody : means and ends." in *Prépublication des Actes du Séminaire Prosodie*, (Aix, octobre 1992) 9-24.

Najim, Z. & Hirst, D.J. (this volume) "Codage prosodique d'un corpus d'arabe littéral lu par un locuteur marocain."

Nicolas, P. (1989) *Amplitude des variations micromélodiques des obstruantes voisées en fonction de la hauteur de la voix*. Mémoire de DEA, Université de Provence.

Roucos & Wilgus (

Scheffers, M.T.M. (1988) "Automatic stylization of F0-contours." in Proceedings of 7th FASE symposium : Speech '88 (Edinburgh),

Taylor, P & Isard, S. (1992) "A new model of intonation for use with speech synthesis and recognition."

“AUTOMATIC MODELLING OF FUNDAMENTAL FREQUENCY.”

Daniel Hirst & Robert Espesser

CNRS, Institut de Phonétique d'Aix

Résumé

On présente un algorithme permettant le codage automatique de la fréquence fondamentale au moyen d'une technique baptisée régression quadratique modale. La sortie de cet algorithme, une séquence de points-cibles <Hz, ms> peut être utilisée en entrée pour la génération de courbes de fréquence fondamentale au moyen d'une fonction spline quadratique.

Abstract

An algorithm for the automatic coding of fundamental frequency is described using a technique called asymmetrical modal quadratic regression. The output of the algorithm, a sequence of target points <Hz, ms>, can be used as input for fundamental frequency synthesis by a quadratic spline function.