

# STYLISATION AND SYMBOLIC CODING OF $F_0$ : A QUANTITATIVE MODEL

Estelle Campione, Emmanuel Flachaire, Daniel Hirst, Jean Véronis

Laboratoire Parole et Langage,  
Université de Provence & CNRS

29 Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1, France

Tel. : +33 4 42 95 36 33, Fax : +33 4 42 59 50 96, E-mail: Jean.Veronis@lpl.univ-aix.fr

## ABSTRACT

This paper presents a reversible model for the stylisation and the symbolic coding of macroprosodic fundamental frequency patterns. Prosodic labels are generated automatically from the speech signal and can be used to regenerate a synthetic  $F_0$  curve which is as close as possible to the original curve. The model has been tested successfully for 20 speakers in French and Italian.

## 1. INTRODUCTION

$F_0$  is often considered as the combination of a macroprosodic component reflecting the speaker's choice of intonation pattern, and a microprosodic component [5] which is entirely dependent on the choice of phonemes in the utterance (lowering of  $F_0$  for voiced obstruents etc.). Numerous studies since the 1960's have attempted to factor out these two components and to extract automatically the relevant macroprosodic information from the speech signal. This extraction can be broken down into two stages :

- *stylisation*, i.e. the replacement of the  $F_0$  curve by a simpler numerical function conserving the original macroprosodic information;
- *symbolic coding*, i.e. the representation by means of an alphabet of symbols, reducing the stylised curve to a sequence of discrete categories.

The first stage is often referred to as *close-copy stylisation* [4] which aims to replace the original  $F_0$  curve by a stylised curve perceptually indistinguishable from the original. The discrete categories of the second stage can be used to re-generate a curve which may be distinguishable from the original one but which is considered by listeners as linguistically equivalent (which De Pijper *op. cit.* calls "standardised perceptual equivalence"). The results presented in this paper suggest that it may be possible to extend the notion of *close-copy* to the symbolic representation, ensuring that no relevant prosodic information is lost in the process of stylisation and symbolic coding: the curve generated from the symbolic coding would then be perceptually indistinguishable from both the stylised curve and from the original curve. This would result in a totally reversible system of analysis: an extremely valuable tool for the automatic coding of large speech corpora.

Stylisation has been the object of a great number of studies ([3], [7], [14], [16]) and it can be said that the technique has been mastered fairly satisfactorily. A number of systems of symbolic coding have also been proposed, but automatisation and reversibility (*close copy*) are far less advanced for the symbolic coding than for the stylisation. In this paper we present a system of stylisation and of symbolic coding which allows the generation of an  $F_0$  curve which is very close to the original curve. The system has been applied successfully to French and to Italian.

## 2. STYLISATION

The method of stylisation used in this study: MOMEL (MODélisation de MELodie) was originally proposed by [7] (see also [8]). Contrary to most other methods of stylisation which use a sequence of straight line segments, MOMEL uses a quadratic spline function (sequence of parabolic segments) resulting in a continuous, smooth curve, without the angles which occur when using straight lines. Unvoiced segments are interpolated so that the resulting curve presents no discontinuities at all. These characteristics of the quadratic spline function are also shared by the more complex stylisation functions used by Fujisaki and colleagues [6].

It has been argued [13] that stylisation by curvilinear functions is not perceptually distinguishable from that using straight-lines. We note however that :

- stylisation by quadratic splines produces a curve which is closer to the original  $F_0$  curve and hence introduces less noise into quantitative studies — in particular in the evaluation of models as in this paper;
- stylisation by quadratic splines produces a macroprosodic contour which is practically identical to the  $F_0$  curves produced on utterances consisting entirely of sonorant segments which are both continuous and smooth.

The quadratic spline functions used for synthesis can be defined by a sequence of target points corresponding to the significant changes of the  $F_0$  curve (zero-crossings of the first derivative).

### 3. SYMBOLIC CODING.

The best-known system for the symbolic coding of intonation at present is ToBI [12] which has been used successfully by numerous researchers for American English. ToBI is a system which is based on an extensive phonological analysis of the intonation system of English and its application to other languages or dialects, while theoretically possible, necessitates a considerable amount of prior research to establish the inventory of prosodic and intonational patterns of the language [10]. ToBI labelling also relies on linguistic judgements made by an expert and is consequently difficult to carry out automatically, although attempts have been made to do this [1]. Finally, the regeneration of an  $F_0$  curve from the ToBI coding is far from obvious.

In our study, we used the symbolic coding system INTSINT (INternational Transcription System for INTonation) described by [9] and [8]. Unlike ToBI which encodes events of a linguistic nature, INTSINT aims to provide a purely formal encoding of the macroprosodic curve. Each target point of the stylised curve is coded by a symbol either as an absolute tone, defined globally with respect to the speakers pitch-range or as a relative tone, defined locally with respect to the immediately neighbouring target-points.

The absolute tones are :

- T (*top*): top of speaker's pitch-range
- B (*bottom*): bottom of speaker's pitch-range
- M (*mid*): initial, mean value.

Relative tones can be coded as :

- U (*up*): target in a rising sequence
- D (*down*): target in a falling sequence
- S (*same*): target not different to preceding target
- H (*higher*): target higher than both immediate neighbours
- L (*lower*): target lower than both immediate neighbours

This system can be thought of as a first degree of abstraction which can be used for the development of more abstract systems such as ToBI. INTSINT has been applied manually to a number of different languages [9]. The automatic and reversible coding poses a number of problems, however. These include :

- determining a threshold beyond which target points are coded as T or B;
- determining the relationship between a target-point and the following point coded as a relative tone H, L, U or D.

### 4. STATISTICAL STUDY

We first studied the characteristics of the distribution of target-points as well as the intervals between successive target-points for a corpus of French and Italian. The corpus chosen was EUROM1 [2] consisting of 35 minutes of speech for French and 54 minutes for Italian. The sentences of the corpus are grouped into continuous passages of 5 sentences and are spoken by 10 speakers per language (5 male, 5 female).

The whole corpus was stylised automatically using the MOMEL algorithm, then checked and corrected manually by experts. Around 5% of the target points were corrected (often minimally). The analysis of the results showed that a certain number of systematic errors (in particular before silent pauses) could be avoided by a slight modification of the algorithm.

The corpus analysed provided 6329 target points for French and 9804 for Italian. The fundamental frequency of the target points was converted to semi-tones (STs). The standard deviation varied from 2.79 to 4.18 STs for the French speakers and from 2.78 to 4.44 STs for the Italian speakers. The mean value of the initial targets for each speaker was very close to the overall mean for all targets for that speaker (mean difference -0.88 STs for French and 0.34 STs for Italian).

The shape of the distribution of target points for each speaker was approximately normal with neither mode nor discontinuity allowing us to set a threshold to distinguish T and B from the other tones. Similarly the distribution of the successive pitch intervals (difference in STs between two successive targets) showed neither mode nor discontinuity (apart from that between rising and falling intervals) allowing such a classification.

A study of linear regressions between successive target points showed no significant correlation for the complete set of targets. A fairly strong correlation was observed however when sequences of 2 points were classified as either rising or falling. Quadratic or exponential transformations did not significantly improve the correlation coefficient.

Finally it was observed that the size of the pitch interval was not significantly correlated with the temporal distance between the two corresponding targets.

### 5. MODEL 1

The first model used a coding obtained in the following way from the stylised  $F_0$  curve:

- the symbol M is assigned to a fictitious target point situated 100 ms before the actual beginning of the utterance;
- target points higher than a threshold  $\tau_T$  are coded T, those lower than a threshold  $\tau_B$  are coded B;

- target-points less than 2.5% of  $\tau_T - \tau_B$  from the previous target are coded S;
- for the remaining target points those corresponding to peaks or valleys are coded H or L respectively, while those in rising or falling sequences are coded U or D.

Different symbolic codings of the corpus were obtained using thresholds  $\tau_T$  and  $\tau_B$  adapted so that the same fixed proportion  $\theta$  of the targets were coded T and B, with  $\theta$  fixed to 5%, 10%, 15%, etc. assuming a normal distribution. Within each language the data for the different speakers was centred and reduced so that they could be pooled.

For each of the values for  $\theta$ , four sets of linear regression coefficients were obtained for each language, allowing the prediction of a target coded U, H, D or L from the value of the preceding target.

In the regeneration:

- targets coded M were assigned the mean value for the speaker;
- targets coded T and B were respectively assigned a value  $F_T$  and  $F_B$  determined by the mean of the values above  $\tau_T$  and below  $\tau_B$  in a normal distribution;
- targets coded S were assigned the same value as the preceding target;
- the value of targets coded U, H, D and L were calculated by applying the regression coefficients to the preceding target.

The precision of the regeneration was measured by the variance of the differences between the original target points and the modelled values. The results for the French data are summarised in Table 1 (results for Italian were similar). Precision is shown by the line *mod*. The line *max* shows the limit of the precision which can be obtained for the given threshold, owing to the fact that the extreme points T and B were assigned a fixed value. This limit would be obtained if all the relative target points were perfectly modelled. The line *HL* represents a variant of the model where the distinction between H and U on the one hand, and between L and D on the other hand, is ignored.

It can be noted that there is quite a substantial residual noise (variance of 0.172 in the best version). Furthermore the optimal thresholds  $\tau_T$  and  $\tau_B$  involve coding 50% of the targets as either T or B which is far from satisfactory, as can be seen on the line *max*, which shows an important degradation of the reachable precision. The distinctions between U and H on the one hand and between D and L on the other hand, are only slightly relevant, for small values of  $\theta$ .

Evaluation by experts showed that while the regenerated curves (using TD-PSOLA) were perceptually close to the original curves (see the notion of "Standardised Percep-

tual Equivalence" in the introduction) they were often distinguishable from the original recordings and could not be considered close-copies. The symbolic coding of Model 1 consequently involved some loss of information.

$\theta$	5%	10%	15%	20%	25%	30%	35%
<i>max</i>	<u>0.011</u>	0.029	0.053	0.083	0.119	0.157	0.198
<i>HL</i>	0.347	0.246	0.202	0.179	<u>0.172</u>	0.183	0.209
<i>mod</i>	0.289	0.212	0.179	0.169	<u>0.168</u>	0.182	0.208

Table 1 : Precision of the re-generated target-points

## 6. MODEL 2

The second model tested combined the idea of relative coding from INTSINT with the idea of pitch levels as described for example by [11]. In this model, target points coded M and S are treated in the same way as in the previous model.

- The thresholds  $\tau_T$  and  $\tau_B$  are fixed so that 5% of the points are coded T and 5% coded B (supposing a normal distribution).  $F_T$  and  $F_B$  are determined as in Model 1.
- Relative targets are coded solely with respect to the previous value as H or L. The distinction between H and U and that between L and D is consequently ignored.
- The central band between  $F_T$  and  $F_B$  is divided into three levels each containing 30% of the points in a normal distribution : G (*grave*), M (*medium*) and A (*acute*). The coding of the relative targets is obtained by associating the relationship to the preceding value (H or L) and the level (G, M or A) in which the target is situated :  $H_G, H_M, H_A, L_G, L_M, L_A$ . Twelve sets of regression coefficients (2 directions and 6 sequences of two levels) were estimated as in the preceding model.

The objective evaluation revealed results which were a considerable improvement on the first model: the residual noise was reduced to 0.081 and 92.93% of the target-points were less than 0.5 ST from the original target. The residual errors concerned essentially the extreme points which were furthest above  $F_T$  or below  $F_B$ . Evaluation by experts showed that differences were perceptible only for the extreme values mentioned above.

Results were similar for Italian (variance of 0.096; 92.86% of target points less than 0.5 ST from the original target). Moreover, the regeneration of the Italian  $F_0$  with the parameters derived from French, even though the language and the speakers were different, showed only a very slight degradation (variance of 0.097). The model consequently appears quite robust and relatively independent of both speakers and language (at least for a

given speaking style and for fairly closely related languages).

## 7. CONCLUSION AND PERSPECTIVES.

The second model examined in this study allows an automatic symbolic coding of  $F_0$  with an excellent reversibility which should prove extremely valuable for the automatic analysis of large speech corpora. This model could be improved in a number of ways — extreme values might be better modelled either by reducing the percentage of Ts and Bs or by adding categories corresponding to extra-high or extra-low values [9]. The number of parameters of Model 2 could no doubt be reduced considerably: one possibility would be to maintain the original distinction of INTSINT between H and U on the one hand and L and D on the other, but to found this distinction not on the local configuration but rather on the size of the pitch interval with respect to the previous target. Further research will be necessary to strike a satisfactory balance between the precision of the model and the number of parameters it involves.

In this study the parameters were evaluated for all speakers pooled within each of the two languages. It would however be interesting to look closer at the inter-speaker variability as well as differences between male and female speakers generally. A further question concerns how far it will be possible to extend the analysis to other (particularly non-romance) languages.

## 8. ACKNOWLEDGEMENTS

The authors thank Robert Espesser for his technical assistance. Robert Espesser is the author of the software used in this study to edit and manipulate speech signals.

## 9. REFERENCES

- [1] Black, A., Hunt, A. (1996). Generating F0 contours from ToBI labels using a linear Regression, *Proc. ICSLP'96*, Philadelphia.
- [2] Chan, D., Fourcin, A., Gibbon, D., Granstrom, B., Hucvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., Zeiliger, J.. (1995). EUROM- A Spoken Language Resource for the EU. *Proc. Eurospeech'95*. Madrid, 1, 867-870.
- [3] D'Alessandro, C., Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception, *Computer Speech and Language*, 9, 257-288.
- [4] De Pijper, JR. (1979). Close-copy stylisation of British English intonation contour, *IPO Annual Progress Report* 14, 66-71.
- [5] Di Cristo, A. Hirst, D.J. (1986). Modelling French micromelody : analysis and synthesis. *Phonetica*, 43, 1/3, 11-30.
- [6] Fujisaki, H., Hirose, K. (1982). Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. *Proc. 13th International Congress of Linguists*, 57-70.
- [7] 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.
- [8] Hirst, D., Di Cristo, A., Espesser, R. (forthcoming). Levels of representation and levels of analysis for the description of intonation systems. In Horne, M. (Ed.), *Prosody: Theory and Experiment*, Dordrecht: Kluwer Academic Publishers.
- [9] Hirst, D., Di Cristo, A. (in press) A survey of intonation systems. in Hirst, D., Di Cristo, A. (Eds) *Intonation Systems: a Survey of Twenty Languages*. Cambridge : Cambridge University Press, 1-44.
- [10] Pierrehumbert, J. (forthcoming). Tonal elements and their alignment. in M. Horne (ed.) *Prosody : Theory and Experiment*. Dordrecht: Kluwer Academic Publishers.
- [11] Rossi, M., Chafcouloff, M. (1972). Les niveaux intonatifs. *Travaux de l'Institut de Phonétique d'Aix*, 1, 167-176.
- [12] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. (1992). ToBI: a standard for labelling English prosody. *Proc. ICSLP'92*, 2, 867-870, Banff, Canada.
- [13] t'Hart, J., (1991). F0 stylization in speech : straight lines versus parabolas, *JASA*, 6, 3368-3370.
- [14] t'Hart, J., Collier, R., Cohen, A. (1990). *A perceptual study of intonation : an experimental-phonetic approach to speech melody*, Cambridge Univ. Press.
- [15] Taylor, P. (1993). Automatic Recognition of Intonation from F0-contours using the Rise/Fall /Connection Model, *Proc. Eurospeech'93*, Berlin, 2, 789-792.
- [16] Taylor, P. (1994). The Rise/Fall/ Connection Model of Intonation, *Speech Communication*, 15:1&2, 169-186.