

Automatic Stylistation and Symbolic Coding of F_0 : Implementations of the INTSINT Model

ESTELLE CAMPIONE, DANIEL HIRST and JEAN VERONIS

0.1 Introduction

F_0 curves are often considered the combination of a macroprosodic component reflecting the speaker's choice of intonation pattern, and a microprosodic component (Di Cristo and Hirst, 1986) 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 (De Pijper, 1979) which replaces the original F_0 curve by a stylised curve that is assumed to be perceptually identical to the original. The discrete categories of the second stage can be used to re-generate a curve which, unlike the perceptually identical version may be distinguishable from the original one but is nonetheless considered by listeners as linguistically equivalent. De Pijper (*op. cit.*) calls this "standardised perceptual equivalence".

Hirst *et al.* (forthcoming) distinguish four distinct levels of representation for prosody. At the most abstract level they assume an underlying phonological representation and at the concrete level a physical (i.e. acoustic

or physiological) representation. Between these two extremes they further distinguish a surface phonological representation constructed from discrete phonological categories and a phonetic representation expressed in terms of continuously variable values. In De Pijper's terms, utterances which have the same surface phonological representation might thus be assumed to be perceptually equivalent, whereas utterances with the same phonetic interpretation would be perceptually identical.

Generating F_0 curves from surface phonological representations obviously leaves quite a lot of unexplained variability in the data. This chapter explores the possibility of enriching surface phonological representations to the point where an utterance synthesised from a symbolic coding might be practically indistinguishable from the original utterance. Stylisation has been the object of a great number of studies and the technique has been mastered fairly satisfactorily. A number of systems of symbolic coding have also been proposed, but automatised and reversibility (close copy) are far less advanced for symbolic coding than for stylisation.

A totally reversible system of analysis would obviously constitute an extremely valuable tool for the automatic coding of large speech corpora. Such tools would be useful both for the study of speech in general and for speech technology, in particular speech recognition and speech synthesis (Véronis *et al.*, 1997, Di Cristo *et al.*, 1997, Véronis *et al.*, forthcoming). The rare prosodically labelled corpora which exist at present (e.g. Ostendorf *et al.*, 1995) have required hand labelling by experts. Besides the laboriousness and difficulty of the task, the subjective nature of such labelling reduces the trustworthiness of the results or requires careful control by counter-experts thus increasing the cost yet more. Automatic techniques removing the need for manual intervention, or at least reducing it to a phase of checking and correcting, would obviously be extremely desirable. Unfortunately there are no automatic robust prosodic transcription systems available even for such widely used systems as ToBI (Silverman *et al.*, 1992), although work in this area is in progress (Black and Hunt, 1996, Dusterhoff *et al.*, 1997).

In this chapter we present six progressively more complex versions of an automatic transcription system based on the INTSINT model described in Hirst *et al.* (forthcoming). The six versions were tested on a corpus of read speech in French and the two most complex versions were subsequently also evaluated for a comparable corpus in Italian (representing readings from 20 different speakers for a total of about 90 minutes of speech). Although there is still some room for improvement, the best two versions provide a very close approximation to the original curves since the standard deviation of

error (on a logarithmic scale) was reduced to less than a semi-tone both for the absolute pitch targets and for the relative pitch intervals.

0.2 Stylisation

In this section we give a brief overview of existing techniques including that which we adopted for this study.

The Instituut voor Perceptie Onderzoek (IPO) originally undertook research on the stylisation of F_0 contours with the aim of developing an intonation model for speech synthesis in Dutch (Cohen and 't Hart, 1965). The IPO approach subsequently evolved into a general theory of intonation structure ('t Hart, Collier, Cohen, 1990), through a procedure of analysis by synthesis. The approach is based on the principle that the simplified F_0 curve must be melodically identical to the original curve ('t Hart and Collier, 1975). In the late 70s, De Pijper (1979) introduced the concept of close-copy stylisation where a subject is asked to compare a recording of an utterance with a synthesised version of the utterance in which the F_0 curve has been replaced by a sequence of straight lines (on a logarithmic scale). A close copy is obtained when subjects are unable to distinguish the synthesised version from the original. For synthesis the piece-wise linear approximations obtained by close-copy stylisation were converted to standard pitch movements chosen from a (language-specific) inventory of "perceptually relevant pitch movements". The resulting output, while usually quite easily distinguishable from the original was nonetheless claimed to be as acceptable as the original and hence "perceptually equivalent" ('t Hart, Collier, Cohen, 1990). This approach, originally developed for the intonation of Dutch was subsequently applied to other languages: English (De Pijper, 1983, Willems, Collier and de Pijper, 1988), German (Adriaens, 1991), Russian (Odé, 1989), French (Beaugendre, 1994) and Indonesian (Odé and Van Heuven, 1994).

More recently, Taylor (1993, 1994) proposed a model analysing an F_0 curve as a linear sequence of three primitive elements: Rise, Fall and Connection which can then be related to a phonological representation. The Rise and Fall are interpreted as piecewise parabolas and are hence equivalent to the quadratic spline presented below. The Connection element is interpreted as a linear transition.

D'Alessandro and Mertens (1995) present a technique for automatic stylisation based on a model of tonal perception which assumes (following House, 1990) that the syllable is the basic perceptual unit for speech. Syllabic pitch-contours are categorised as dynamic or static depending on

the presence of a perceptible (and possibly complex) pitch movement. The F_0 contour is transformed into a sequence of tonal segments which are either static or dynamic as a function of a glissando threshold which varies with the duration of the syllable.

The method of stylisation used in this study: MOMEL (MODélisation de MELodie) was originally proposed by Hirst (1980, 1983) and automated by Hirst and Espesser (1993) (see description in Appendix I). Contrary to many 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 (Fujisaki and Hirose, 1982) as the continuation of earlier work by Öhman (1967). This model is based on the hypothesis that continuously varying F_0 curves are the result of a sequence of discrete commands produced by the speaker. Fujisaki distinguishes phrase commands and accent commands modelled respectively as an impulse-like command and a step-command.

It has been argued ('t Hart, 1991) 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 smooth¹.

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 of the first derivative²). Appendix I summarises the algorithm.

0.3 Symbolic coding

The most widely used system for the symbolic coding of intonation at present is ToBI (Silverman *et al.*, 1992) which has been used successfully by numerous researchers for American English, although it has been

criticised for falling unsatisfactorily between a phonetic and a phonological system (see for example Nolan and Grabe, 1997). 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, would necessitate a considerable amount of prior research to establish the inventory of intonation patterns of the language (Pierrehumbert, forthcoming). ToBI labelling also relies on linguistic judgements made by experts and is consequently difficult to carry out automatically, although attempts have been made to do this (Wightman and Ostendorf 1992; Ostendorf and Ross, 1997). Finally, the regeneration of an F_0 curve from the ToBI coding is far from obvious.

In our study, we use a symbolic coding system INTSINT (INternational Transcription System for INTonation) proposed by Hirst and Di Cristo. (1998) and Hirst *et al.* (forthcoming) and which has been used for the manual transcription of intonation patterns in a number of languages (see different chapters in Hirst and Di Cristo, 1998). 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. Relative tones can further be subdivided into iterative and non-iterative categories where it is assumed that iterative tones can be followed by the same tone whereas non-iterative tones cannot. Within each category, positive, negative and neutral values are defined, which thus give a total of nine possible tone categories. Of these, it is assumed that one logical possibility, that of an iterative neutral tone, does not in fact occur so that the following eight possibilities are actually used (Table 0.1).

Table 0.1. Orthographic and iconic symbols for the INTSINT coding system. The letters stand for Top, Mid, Bottom, Higher, Same, Lower, Upstepped and Downstepped respectively.

		<i>Positive</i>	<i>Neutral</i>	<i>Negative</i>
<i>ABSOLUTE</i>		T [↑]	M [⇒]	B [↓]
<i>RELATIVE</i>	<i>Non-Iterative</i>	H [↑]	S [→]	L [↓]
	<i>Iterative</i>	U [<]	•	D [>]

This system can be thought of as a surface phonological system, something along the lines of the International Phonetic Alphabet for the transcription of segmental phonology. It can thus be seen as a first degree of

abstraction which can be used for the automatic extraction of "linguistic-like" representations from spoken corpora which could provide the data necessary for the development of more abstract phonological systems such as ToBI.

The automatic and reversible coding using INTSINT poses a number of problems, however. Nicolas and Hirst (1995) showed, for example, that for the coding of continuous texts, considerable improvement was obtained when the values of the extreme tones T and B were varied to take into account paragraph level effects. In this chapter we concentrate on the coding of the relative tones for which the two basic problems are:

- determining a threshold to separate absolute tones T, B from relative tones H, S, L, U, D;
- determining the optimal criteria for distinguishing iterative and non-iterative tones

The distinction between iterative and non-iterative tones can be based on two rather different criteria. The first distinction is purely configurational in that basically H and L are respectively peaks and valleys whereas D and U are plateaus in rising/falling sequences of tone. If this were the only distinction, the coding would be redundant and U and D could be treated as allophones of H and L. There is, however, also a second scalar distinction between the two categories, since in general H and L are assumed to correspond to larger frequency intervals than U and D.

In the rest of this paper we present results on the evaluation of six different implementations of the INTSINT model. In particular we separate out the configurational and the size criteria in order to evaluate the relative importance of each. We also investigate the possibility of extending the scalar values in order to come nearer to a close-copy stylisation³.

0.4 Common properties of the six implementations

As mentioned above the six implementations we tested are all based on the automatic stylisation technique MOMEL. In all the implementations, two absolute symbols T and B are used to code extreme pitch values. The symbol S is used to code target points which are not significantly different from the preceding point.

- target points higher than a threshold τ_T are coded T, those below a threshold τ_B are coded B ;

Automatic Stylistation and Symbolic Coding of F_0

- target points less than 2.5% (on a log scale) from the preceding point are coded S.

The thresholds t_T and t_B are chosen so that 5% of the target points are coded T and another 5% are coded B, assuming a normal distribution of the values of target-points on a logarithmic scale, which a prior analysis (see below) showed to be a satisfactory approximation.

The resynthesis of the target points takes as a starting value a value situated before the beginning of each utterance coded M with an F_0 value equal to that of the overall mean of the target points. Subsequent points are generated in the following way:

- points coded T and B are assigned the values F_T and F_B corresponding to the mean of the target points respectively above τ_T and below τ_B assuming a normal distribution.
- points coded S are assigned the same value as the preceding target point⁴;
- the value of the other target points are calculated by linear regression from the value of the preceding target point using regression coefficients estimated separately for each symbol on the z-transformed values for all speakers pooled for each language.

It should be noted that while there is no specific "downdrift" or "declination" component in this model, the typical lowering of a sequence of successive H and L targets observed in natural speech is an automatic consequence of applying the same regression coefficients iteratively. Thus sequences such as [M T L H L H L H...] or [M T D D D ...] will produced target values which decline towards asymptotic values (cf Liberman and Pierrehumbert, 1986, Hirst *et al.*, forthcoming).

0.5 Specific characteristics of the different implementations

0.5.1 Version HL

In a purely configurational interpretation, the distinction between iterative and non-iterative tones is treated as simply allophonic. The first implementation we tested is a minimal one conflating the two categories and assuming only two symbols, H and L (in addition to T, B and S which as mentioned above are common to all implementations) (Figure 0.1):

- H : rising ;
- L : falling.



Figure 0.1. Examples of coding with version *HL*.

0.5.2 Version Config

The second implementation is designed to test the possibility that the difference between iterative and non-iterative values, while allophonic, nevertheless requires separate regression coefficients. Target points are consequently coded as H, L, U or D according to configuration (Figure 0.2):

- H : peak ;
- U : raised plateau (upstepped) ;
- D : lowered plateau (downstepped) ;
- L : valley.

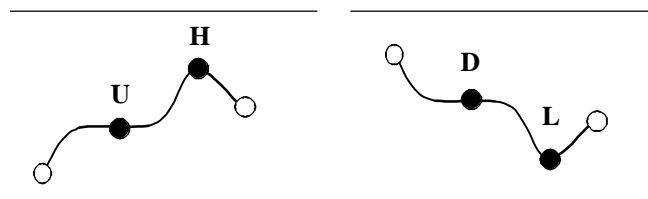


Figure 0.2. Examples of coding with version *Config*.

0.5.3 Version Mixed

This version is an implementation of the algorithm described in Hirst *et al.* (forthcoming). This version assumes that the distinction between iterative and non-iterative categories is based both on configuration and on size. Thus

plateaus in rising or falling sequences will necessarily be coded U and D respectively as in the *Config* version, whereas peaks will be coded either H or U and valleys will be coded either L or D depending on the size of the pitch interval with respect to the preceding target (Figure 0.3).

- H: peak AND interval greater than a threshold α above the previous target point;
- U: raised plateau (upstepped) OR interval less than a threshold α above the previous target point;
- D: lower plateau (downstepped) OR interval less than a threshold α below the previous target point;
- L: valley AND interval greater than a threshold α below the previous target point.

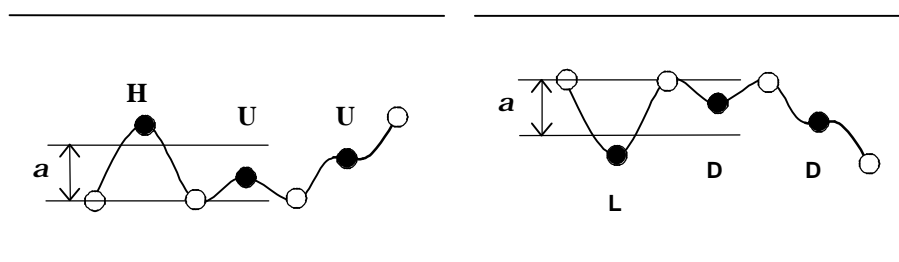


Figure 0.3. Examples of coding with version *Mixed*.

0.5.4 Version *Ampli2*

The implementation *Ampli2* assumes that the distinction between iterative and non-iterative tones is purely scalar, i.e. that H and L represent larger pitch intervals with respect to the preceding target point than do U and D respectively (Figure 0.4).

- H_1 : interval less than a threshold α above the previous target point;
- H_2 : interval greater than a threshold α above the previous target point;
- L_1 : interval less than a threshold α below the previous target point;
- L_2 : interval greater than a threshold α below the previous target point;

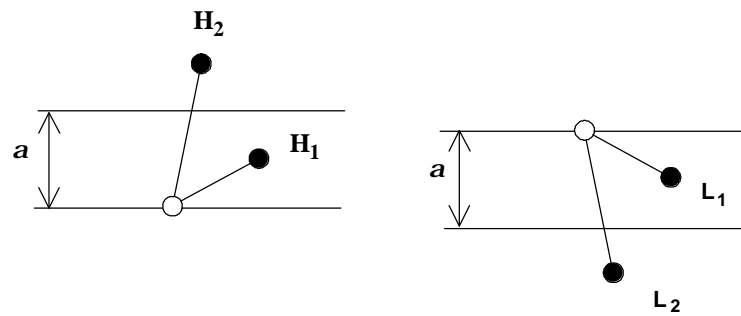


Figure 0.4. Examples of coding with version *Ampli2*.

0.5.5 Version *Ampli3*

The implementation *Ampli3* is based on the same principle as version *Ampli2* except that the scalar aspect is extended to distinguish three degrees of size defined by two thresholds α_1 and α_2 (Figure 0.5). This allows us to define six relative pitch levels:

- H_3 : interval greater than α_2 above the previous target point;
- H_2 : interval between α_1 and α_2 above the previous target point;
- H_1 : interval less than α_1 above the previous target point;
- L_1 : interval less than α_1 below the previous target point;
- L_2 : interval between α_1 and α_2 below the previous target point;
- L_3 : interval greater than α_2 below the previous target point.

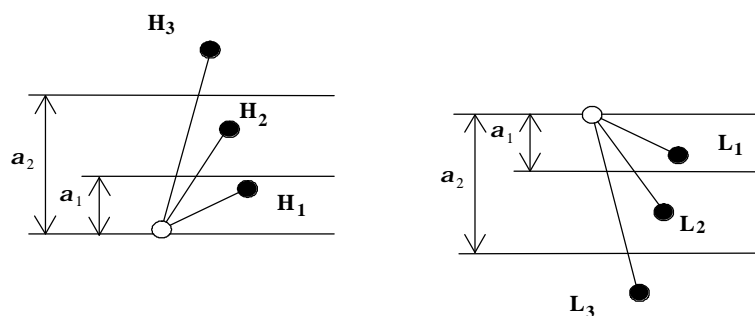


Figure 0.5. Examples of coding with version *Ampli3*.

0.5.6 Version Levels

The final implementation we tested explores a rather different principle, following earlier research by Rossi and Chafcouloff (1972). The central region between the two thresholds t_T and t_B is divided into three bands: G(rave), M(edium), A(cute), each corresponding to one third of the target points assuming a normal distribution. The coding of the target points takes into account both the direction with respect to the preceding target point and the band in which the targets are situated (Figure 0.6) :

- H_A : higher than previous target and finishing in the acute band ;
- H_M : higher than previous target and finishing in the medium band;
- H_G : higher than previous target and finishing in the grave band;
- L_A : lower than previous target and finishing in the acute band;
- L_M : lower than previous target and finishing in the medium band;
- L_G : lower than previous target and finishing in the grave band.

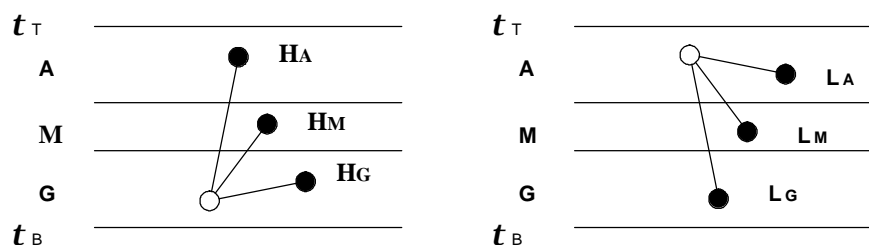


Figure 0.6. Examples of coding with version *Levels*.

0.6 Evaluation

0.6.1 Corpus

The corpus used for this study is part of the multilingual speech corpus EUROM1 which was defined and produced as a deliverable of the European ESPRIT project SAM (Multi-lingual Speech Input/Output Assessment,

Methodology and Standardisation, Chan *et al.*, 1995). We used the French and the Italian versions of the corpus, specifically the part consisting of 40 continuous passages read by ten speakers (5 male and 5 female) in each language. Each passage comprises 5 semantically linked sentences; the French and Italian versions are fairly free translations from the English version.

The following is a sample passage in English, French and Italian :

Please take a request for an early-morning taxi. Mr Spencer of Chestnut Drive wishes to be at Heathrow terminal 4 by 6.15 a.m. His flight's not leaving till 7.50 but he has to arrange for excess baggage. Mark it as top priority and ensure punctuality. He expects you at 6.15sharp.

Je voudrais commander un taxi pour demain matin de très bonne heure. C'est pour monsieur Durand, 4 rue des Châtaigniers. Il veut être à l'aéroport de Marignane, au départ des lignes intérieures, avant six heures et quart. Son avion décolle seulement à huit heures moins dix mais il a besoin de faire enregistrer un excédent de bagages. Il faudra absolument être à l'heure. Je compte sur vous. Il vous attendra en bas de chez lui à cinq heures et demie précises.

Vorrei prenotare un taxi per domani mattina presto per il dottor Rossi all'Hotel Sheraton. Deve essere all'aeroporto per le sei e trenta. Può arrivare per le sei? E' richiesta la massima puntualita'. Può lasciarmi la sua sigla per favore?

0.6.1.1 The French corpus

The French corpus consists of forty passages of 5 sentences - a total of 200 different sentences. The passages are in four groups of 10, each of which is read by ten speakers so that:

- each of the ten speakers read one group of ten passages;
- the first two groups were read by three speakers;
- the second two groups were read by two speakers.

In all, 100 passages (500 sentences) were recorded, totalling 36 minutes 51 seconds of speech. Each sentence was read by either two or three speakers and the duration per speaker ranged from 3 minutes 28 seconds to

4 minutes 37 seconds.

0.6.1.2 The Italian corpus

The Italian corpus also contains forty passages of five sentences, a total of 200 sentences. The passages are in eight groups of five and were recorded by ten speakers so that:

- each of the ten speakers recorded three groups of five passages;
- the first six groups were recorded by four speakers ;
- the last two groups were recorded by three speakers.

In all, 150 passages (750 sentences) were recorded totalling 54 minutes 31 seconds of speech. Each sentence was recorded by either 3 or 4 speakers and the duration per speaker varied from 5 minutes 2 seconds to 7 minutes 11 seconds.

0.6.1.3 Stylistation and correction

Once the entire corpus had been stylised using the MOMEL algorithm, manual corrections were made by experts using a minimalist strategy, making corrections only when otherwise there was an audible difference between the original recording and a recording produced from the stylised F_0 by means of PSOLA resynthesis (Hamon *et al.*, 1989). Apart from a slight distortion due to the lack of microprosodic effects in the stylised curves, the corrected versions were considered by the experts perceptually identical to the original recordings. Two different methods (qualitative and quantitative) were used to analyse the results (Campione and Véronis, 1998a and b).

0.6.1.4 Statistical analysis

We first studied the characteristics of the distribution of target-points as well as the intervals between successive target-points for the French and Italian corpora (see a detailed analysis in Campione and Véronis, 1998c). 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 of the passages 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 providing an

objective criterion to set a threshold distinguishing 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.

While linear regressions between successive target points obviously showed no significant autocorrelation for the complete set of targets, a fairly strong correlation was observed 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.

The six implementations of the model were first evaluated on the French corpus. Subsequently, the best two versions were also tested for Italian.

0.6.2 Method and Result

The curves resynthesised from the six different versions of the automatic coding were compared with the stylised curves corrected by the experts. As has often been noted in the literature, measuring the similarity of two F_0 curves is not an easy task and ideally should make use of perceptual tests which are rarely practical for large corpora. We used the following measures:

- e_{abs} : standard deviation of error (SDE) of the fundamental frequency of the target points (in STs) ;
- e_{rel} : standard deviation of error (SDE) of the interval between two successive target points (in STs) ;
- p_{abs} : the percentage of target points less than 2 STs from the original target ;
- p_{rel} : the percentage of intervals less than 2 STs from the original interval.

The results are summarised in (Figure 0.7) and (Figure 0.8). For French, the value of e_{abs} is minimal for the version *Levels* (0.96 STs), that of e_{rel} is

Automatic Stylistation and Symbolic Coding of F_0

minimal for the version *Ampli3* (1.09 STs). The percentage p_{abs} is maximal for the version *Levels* (96%) and the percentage p_{rel} is maximal for the version *Ampli3* (93.1%).

Similar results were obtained for the Italian corpus on the two versions *Levels* and *Ampli3* which were the only ones tested (Table 0.2). The fit for the data was globally slightly worse than for the French speakers. This was probably due to a greater variability of extreme values for the Italian speakers, particularly the female speakers (Campione, 1997).

Appendix II gives an example of regeneration of the same fragment using the six versions.

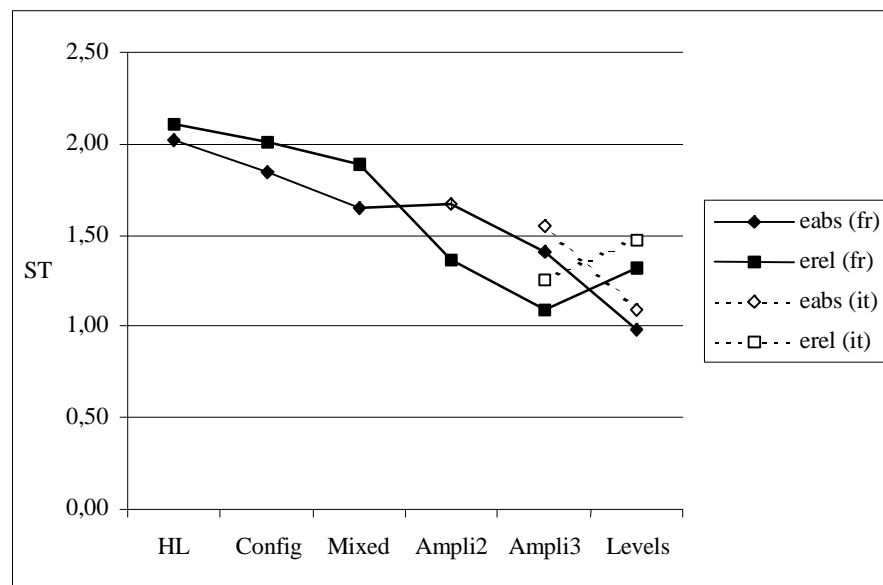


Figure 0.7. Standard deviation of error (SDE).

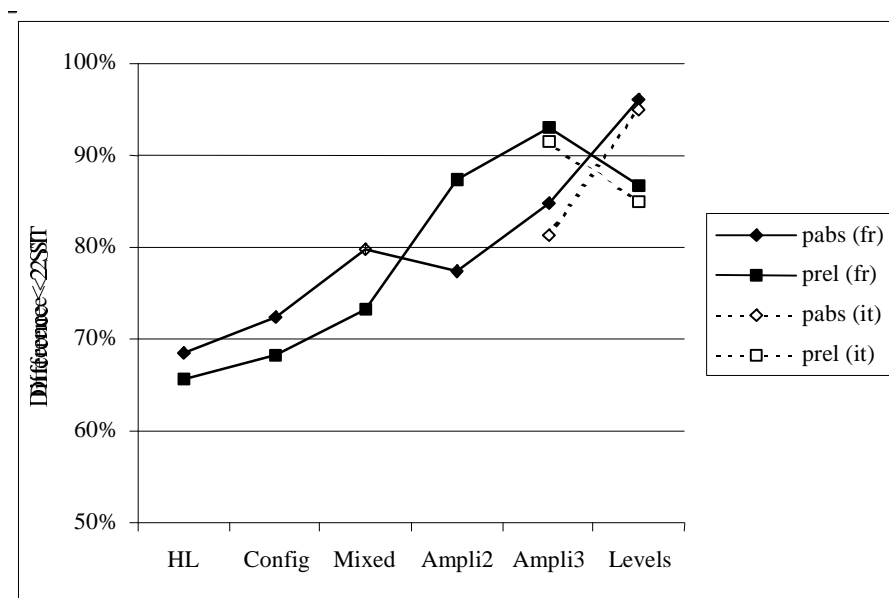


Figure 0.8. Percentage of intervals less than 2 STs from the original data.

Table 0.2. Comparison of result for the best two versions (*Ampli3* and *Levels*) as applied to French and Italian.

	<i>Ampli3</i>		<i>Levels</i>	
	<i>Fr</i>	<i>It</i>	<i>Fr</i>	<i>It</i>
e_{abs} (ST)	1.41	1.55	0.98	1.09
e_{rel} (ST)	1.09	1.25	1.32	1.47
p_{abs} (%)	84.9	81.4	96.0	95.0
p_{rel} (%)	93.1	91.6	86.8	84.9

0.7 Discussion

As was to be expected, the precision of the coding generally increased with the number of symbols used. We note however the following points:

- The *HL* version is clearly insufficient for any practical use.
- For the same number of symbols, the *Mixed* version was superior to the *Config* version for both absolute and relative measures. In turn the *Ampli2* was superior to the *Mixed* version, but only for the relative measures, the absolute measures being slightly worse. However, it is likely that a better fit on relative movements is more desirable in

perceptual terms than a perfect fit of absolute target values. A finer discrimination between these two versions (as well as between *Ampli3* and *Levels*) would need to take into account more results from subjective perception tests in order to determine whether relative intervals or absolute pitch levels are more crucial for the global perception of pitch curves. This result indicates that configuration alone is perhaps not sufficient to define the difference between iterative and non-iterative tones. On the other hand the fact that the *Ampli2* model performed better than the *Mixed* model for relative measures suggests that the distinction between small and larger intervals seems to be relevant not just for peaks and valleys but also for plateaus in rising and falling sequences. Further studies will be necessary in order to determine what linguistic or paralinguistic functions such a distinction embodies.

- The best two versions, *Ampli3* and *Levels*, for the same number of symbols have opposite results in terms of relative and absolute measures. *Ampli3* provides a better fit for relative pitch intervals whereas *Levels* provides a better fit for absolute target values. Similar results were obtained for both French and Italian which suggests that this difference might be relatively language independent. It should be noted that while both *Ampli3* and *Levels* require six relative symbols for the symbolic coding, *Ampli3* used only six sets of regression coefficients whereas *Levels* used 12 since the band of the preceding target was treated as conditioning an allophonic realisation of the symbol. Thus separate coefficients were calculated for e.g. H_A when preceded by a target in the Acute, Medium or Grave band.

A number of other possibilities for improvement remain to be explored. The scalar factor introduced into the *Ampli2* and *Ampli3* models might be increased arbitrarily until some global criteria for fitting was satisfied. There is also, as we mentioned above, room for improvement in the fitting of the extreme values. It seems, nonetheless that the model as it stands is capable of providing a satisfactory model for the behaviour of F_0 target points in the type of data we have been looking at for French and Italian. Adding a scalar dimension to the symbolic coding provides a means of adapting the model with arbitrary precision to a given speaker's pronunciation of an utterance. This would allow us to derive both a perceptually equivalent coding (Basic INTSINT) and a close-copy coding (Scalar INTSINT) for the same utterance, which might prove to be a valuable distinction for the linguistic and paralinguistic interpretation of utterances. Finally a great deal of work remains to be done concerning the relationship between the “linguistic-like”

representations which the algorithms described here allow us to derive automatically and the perceptual and linguistic or paralinguistic interpretation of such representations.

Notes

1. As pointed out by one anonymous reviewer, even sonorant consonants like /m/ may contain some deviations from the macroprosodic curve (often restricted though to one or two periods).
2. Or values very close to zero.
3. One anonymous reviewer has noted that this introduction of scalar values means that the resulting transcriptions are neither phonetic nor phonological but somewhere in between. We agree with this, but feel that the cut-off point between phonological and phonetic representations remains a question for empirical investigation. In our view scalar transcriptions of the type we present here could provide precisely the tool needed for work in this area.
4. Target-points coded S could equally well be assigned a slightly lower value than preceding ones (as in Hirst *et al.* forthcoming) introducing a declination effect.

Acknowledgements

The authors would like to thank Robert Espesser for his technical assistance, Corine Astésano and Fabienne Courtois for their help with the correction of the MOMEL stylisation and Emmanuel Flachaire for his help with the data analysis. Our thoughts go in particular to the memory of Fabienne Courtois who, tragically, met a fatal accident while this work was in progress.

References

- Aasa, A., Stangert, E. 1996. Prosodic Analysis of Swedish within the Multext project, *Fonetik 96*, Nösslingen.
- Adriaens, L.M.H. 1991. *Ein Modell deutscher Intonation : eine experimentell – phonetische Untersuchung nach den perzeptiv relevanten Grundfrequenzänderungen in vorgelesenem Text*. PhD thesis, Eindhoven University of Technology.
- Astésano, C., Espesser, R., Hirst, D. 1997. Stylisation automatique de la fréquence fondamentale : une évaluation multilingue. *Actes du 4ème Congrès Français d'Acoustique*, Marseille, 441-444.
- Beaugendre, F. 1994. *Une étude perceptive de l'intonation du français*. Thèse d'Etat, Université de Paris XI.

Automatic Stylistation and Symbolic Coding of F_0

- Black, A., Hunt, A. 1996. Generating F_0 contours from ToBI labels using a linear regression. In *Proceedings of ICSLP'96*, Philadelphia.
- Campione, E. 1997. *Stylistation et codage symbolique de l'intonation: une étude statistique*. Mémoire de DEA, Université de Provence, Aix-en-Provence.
- Campione, E., Véronis, J. 1998a. Une évaluation de l'algorithme de stylistation automatique MOMEL. *Travaux de l'Institut de Phonétique d'Aix en Provence*, (in press).
- Campione, E., Véronis, J. 1998b. A multilingual prosodic database. In *Proceedings of ICSLP'98*, Sydney.
- Campione, E., Véronis, J. 1998c. A statistical study of pitch target points in five languages. In *Proceedings of ICSLP'98*, Sydney.
- 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. EUROM1 - A Spoken Language Resource for the EU. In *Proceedings of Eurospeech'95*, Madrid, 1, 867-870.
- Cohen, A., 't Hart, J. 1965. Perceptual analysis of intonation pattern. *Actes du 5ème Congrès International d'Acoustique*, Liège, 1-4.
- D'Alessandro, C., Mertens, P. 1995. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, 9, 257-288.
- De Pijper, J.R. 1979. Close-copy stylistation of British English intonation contour. *IPO Annual Progress Report*, 14, 66-71.
- De Pijper, J.R. 1983. *Modelling British English Intonation*. (Netherlands Phonetic Archives 3) Dordrecht : Foris.
- Di Cristo, A., Di Cristo, P., Véronis, J. 1997. A metrical model of rhythm and intonation for French text-to-speech synthesis. In *Proceedings of an ESCA Tutorial and Research Workshop on Intonation : Theory, Models and Applications*, September 1997, Athens, 83-86.
- Di Cristo, A., Hirst, D.J. 1986. Modelling French micromelody: analysis and synthesis. *Phonetica*, 43, 1/3, 11-30.
- Dusterhoff, K., Black, A. 1997. Generating F_0 Contours for Speech Synthesis Using the Tilt Intonation Theory. In *Proceedings of an ESCA Tutorial and Research Workshop on Intonation : Theory, Models and Applications*, September 1997, Athens, 107-110.
- Fujisaki, H., Hirose, K. 1982. Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. In *Proceedings of 13th International Congress of Linguists*, 57-70.
- Hamon, C., Moulines, E., Charpentier, F. 1989. A diphone system based on time-domain prosodic modifications of speech. In *Proceedings of ICASSP'89*, 238-241.
- 't Hart, J., 1991. F_0 stylization in speech : straight lines versus parabolas. *Journal of the Acoustical Society of America*, 6, 3368-3370.
- 't Hart, J., Collier, R. 1975. Integrating different levels of intonation analysis. *Journal of*

- Phonetics*, 3, 235-255.
- 't Hart, J., Collier, R., Cohen, A. 1990. *A perceptual study of intonation : an experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.
- 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 A. and Ladd R. (Eds), *Prosody: Models and Measurements*. Berlin, Springer. 93-109
- Hirst, D.J., Di Cristo, A. 1998. A survey of intonation systems. In Hirst, D.J., Di Cristo, A. (Eds), *Intonation Systems: a Survey of Twenty Languages*. Cambridge : Cambridge University Press, 1-44 (in press).
- Hirst, D.J., Di Cristo, A., Espesser, R. 1998. 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 (forthcoming).
- Hirst, D.J., 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.
- House, D. 1990. *Tonal Perception in Speech*. Lund: Lund University Press.
- Lieberman, M., Pierrehumbert, J. 1984. Intonational invariance under changes in pitch-range and length. In M. Aranoff & I. Sag (eds.), *Language Sound Structure. Studies in Phonology Presented to Morris Halle*, 157-233. Cambridge, Mass.: MIT Press.
- Nicolas, P., Hirst, D.J. 1995. Symbolic coding of higher level characteristics of fundamental frequency curves. In *Proceedings of the 4th European Conference on Speech Communication and Technology* (Madrid 1995).
- Nolan, F., Grabe, E. 1997. Can 'ToBI' transcribe intonational variations in British English?. In *Proceedings of ESCA Workshop Intonation: Theory, Models and Applications*, September 18-20, 1997, Athens (Greece), 259-262.
- Odé, C. 1989. *Russian intonation: a perceptual description*. Amsterdam : Rodopi.
- Odé, C., Van Heuven, V. J. 1994. *Experimental studies of Indonesian prosody*. Department of Languages and Oceania, University of Leiden.
- Öhman, S. 1967. Word and sentence intonation: A quantitative model. *KTH Quarterly Progress and Report*, 2, 25-54.
- Ostendorf, M., Ross, K. 1997. A multi-level model for recognition of intonation labels. In Sagisaka Y., Campbell N. and Higuchi (Eds), *Computing Prosody*. Berlin: Springer, 291-308.
- Ostendorf, M.F., Price P. J., Shattuck-Hufnagel S. 1995. The Boston University Radio News Corpus. *Technical Report No. ECS-95-001*, Boston University.
- Pierrehumbert, J. 1998. Tonal elements and their alignment. In Horne M. (ed.) *Prosody : Theory and Experiment*. Dordrecht: Kluwer Academic Publishers (forthcoming).
- Rossi, M., Chafcouloff, M. 1972. Les niveaux intonatifs. *Travaux de l'Institut de*

Automatic Stylistation and Symbolic Coding of F_0

Phonétique d'Aix, 1,167-176.

- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP'92*, 2, 867-870, Banff, Canada.
- Taylor, P. 1993. Automatic Recognition of Intonation from F_0 contours using the Rise/Fall /Connection Model. In *Proceedings of Eurospeech'93*, Berlin, 2, 789-792.
- Taylor, P. 1994. The Rise/Fall/ Connection Model of Intonation. *Speech Communication*, 15:1&2, 169-186.
- Véronis, J., Di Cristo, P., Courtois, F., Lagrue, B. 1997. A stochastic model of intonation for French text-to-speech synthesis. In *Proceedings of Eurospeech'97*, Rhodes (Greece), September 1997.
- Véronis, J., DiCristo, P., Courtois, F., Chaumette, C. 1998. A stochastic model of intonation for Text-to-Speech Synthesis, *Speech Communication*, forthcoming.
- Véronis, J., Hirst, D., Ide, N. 1994. NL and speech in the MULTEXT project. *AAAI'94 Workshop on Integration of Natural Language and Speech*, Seattle, 72-78.
- Wightman, C.W. Ostendorf, M. 1992. Automatic recognition of intonational features. In *Proceedings of ICASSP'92*, I, 221-224.
- Willems, N., Collier R., De Pijper, J. 1988. A Synthesis scheme for British English Intonation. *Journal of the Acoustical Society of America*, 1984, 1250-1260.

Appendix I: The MOMEL Algorithm

Description [From Hirst & Espesser (1993), Hirst *et al.* (forthcoming)].

The algorithm comprises four stages. After a preliminary pre-processing of F_0 (stage 1), which eliminates aberrant values after voiceless sections, the central part of the algorithm (stage 2) consists in estimating target-candidates by a technique called *asymmetrical modal quadratic regression*. This stage works on the assumption that all relevant microprosodic effects consist in a lowering of the values of the underlying macroprosodic curve. The modal regression is applied within a moving window providing an optimal target for the F_0 values within the window. The next stage (stage 2) partitions the candidate targets and the final stage (stage 3) reduces the targets of each partition to a single candidate.

Stage 1: *Pre-processing of F_0 .* All values 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 (which are often dubious) at the onset of voicing..

Stage2: *Estimation of target-candidates.* This consists in three steps which are followed iteratively for each instant x .

- Within an analysis window of length A (typically 300ms) centred on x , values of F_0 , (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. The threshold $hzmin$ is a constant set to 50 Hz and the adaptive threshold $hzmax$ is set to the mean of the top 5% of the F_0 values of the sequence multiplied by 1.3.
- A quadratic regression is applied within the window to all non-neutralised values. All values of F_0 which are more than a distance Δ below the value of F_0 estimated by the regression are neutralised (typical value of Δ fixed at 5%). This step is re-iterated until no new values are neutralised.
- For each instant x , a target point $\langle t, h \rangle$ is calculated from the regression equation:

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

where $t = -b/(2c)$ and $h = a + bt + ct^2$

Automatic Stylisation and Symbolic Coding of F_0

- This target point corresponds to the extremum (maximum or minimum) of the corresponding parabola (Figure 0.9).

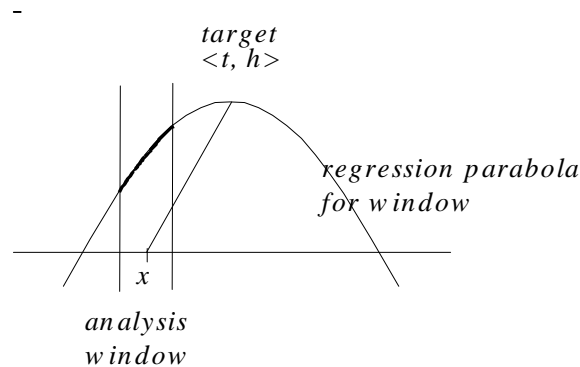


Figure 0.9. Calculation of a local target point.

If h is less than $hzmin$ or greater than $hzmax$, then t and h are treated as missing values.

Steps b) and c) are repeated for each instant x , resulting in one estimated target point $\langle t, h \rangle$ (or a missing value) for each original value of F_0 as in Figure 0.10 where the grey lines connect the value of x to the value of $\langle t, h \rangle$ for each window.

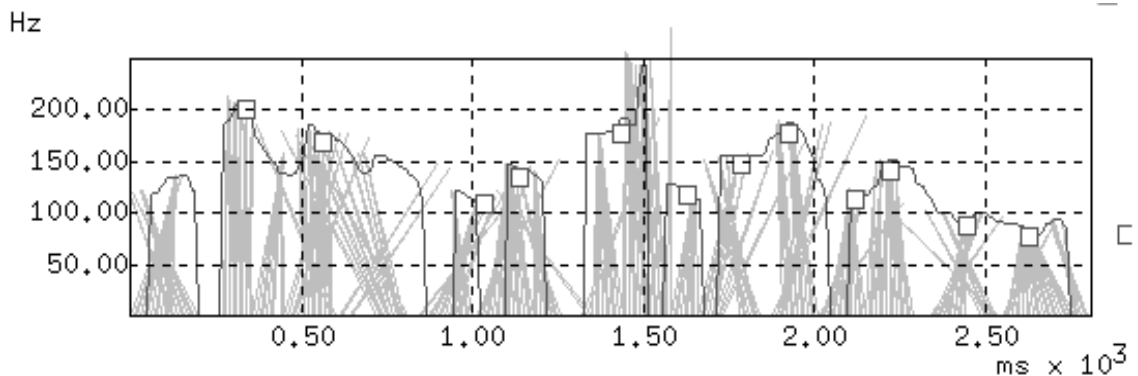


Figure 0.10. Estimation of candidate target point (grey lines) and final targets (white squares). The grey lines connect the centre of the moving window to the extremum of the parabola estimated for that window.

Stage 3: Partitioning of target candidates. The sequence of target candidates is partitioned by means of another moving window R (typically 200 ms) which is divided into two halves, left and right. The partition

algorithm seeks values where there is a maximum difference between the targets in the left and right halves of the window. Specifically, a partition boundary is inserted when the difference between the average weighted values of t and h in the left and right halves of the window corresponds to a local maximum which is greater than a threshold (set to the mean distance between left and right halves for all windows).

Stage 4: Reduction of candidates. Within each segment of the partition, outlying candidates more than one standard deviation from the 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 (Figure 0.11).

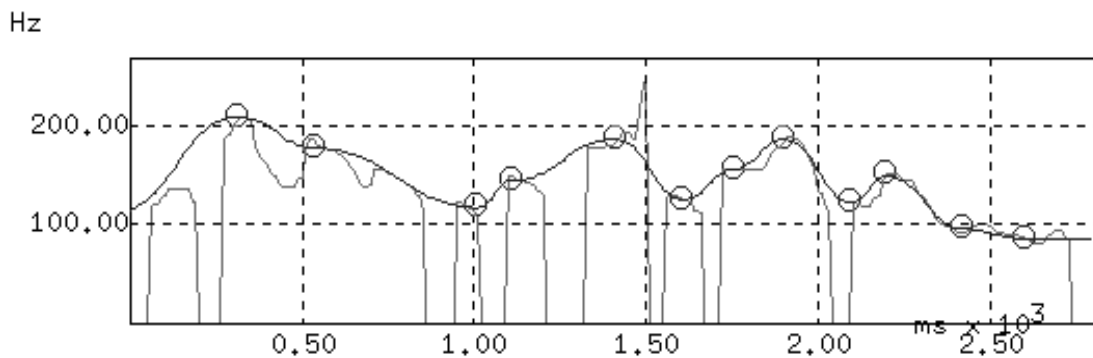


Figure 0.11. F_0 and quadratic spline curve obtained from MOMEL algorithm.

Estimation of parameters and evaluation

The three parameters used by the algorithm (the analysis window $[A]$, the distance threshold $[D]$ and the reduction window $[R]$), were estimated from a small corpus consisting of two sentences containing all the stops and fricatives of French and pronounced by 10 subjects (5 male and 5 female). For the different values of the parameters a subjective evaluation was carried out consisting of a visual and auditory comparison between the original signal and a signal generated by PSOLA resynthesis from the stylised curve. This was completed by an objective analysis consisting of a mean distance between the original and the stylised curve. The optimal values were as follows: $A = 300$ ms ; $D = 5\%$; $R = 200$ ms.

These optimised parameters were subsequently used for the stylisation of other corpora (Hirst and Espesser, 1993 ; Hirst, Di Cristo and Espesser, forthcoming). The results showed that the percentage error (missing or erroneous targets), while slightly higher than for the first corpus was at a relatively reasonable level of 5%. The errors were moreover systematically of two or three different types, in particular missing targets in transitions

from voiced to voiceless segments of speech, which suggests that an improved algorithm could probably eliminate the majority of them.

The stylisation technique has also been applied to a number of other languages (English, German, Arabic, Spanish, French, Italian, Swedish) in the framework of the European MULTTEXT project (Véronis *et al.*, 1994), with comparable results. Astésano *et al.* (1997) describe an evaluation of the algorithm for five languages (English, French, German, Spanish, Swedish) for a duration of 3 hours 45 minutes of speech (see also Aasa et Strangert, 1996 for Swedish). It was also evaluated in depth on French and Italian by Campione (1997). All these studies conclude that MOMEL is an efficient and robust technique for the representation of relevant information of F_0 curves, at least for the languages studied so far with a fairly low error rate of around 5%.

Appendix II. Examples of regeneration with the various versions

Figure 0.12 shows the stylised curve regenerated with the six different versions of the model, along with the original (dotted line). The fragment is "Pourriez-vous m'indiquer à quelle heure est la correspondance à Valence? Si je dois partir avant midi de Marseille, j'aimerais savoir s'il y aura un wagon restaurant." (female speaker). (Could you tell me at what time there is a connection at Valence? If I have to leave Marseilles before midday I'd like to know if there is a dining-car.).

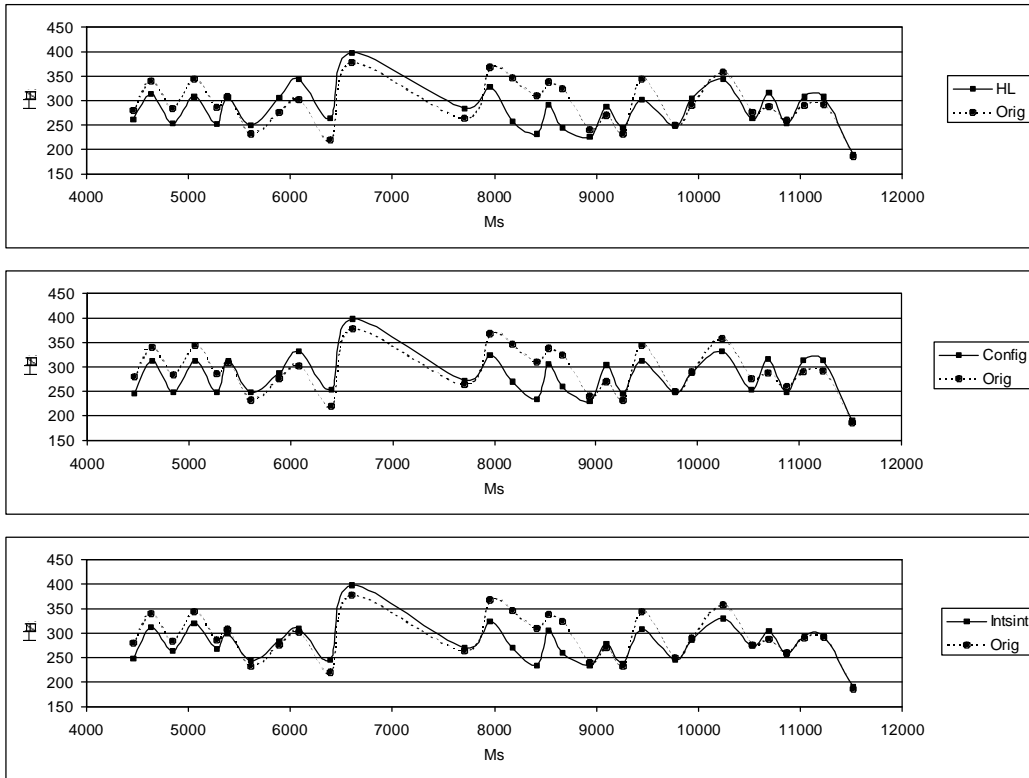


Figure 0.12 (part a). Regeneration with models *HL*, *Config*, *Mixed*

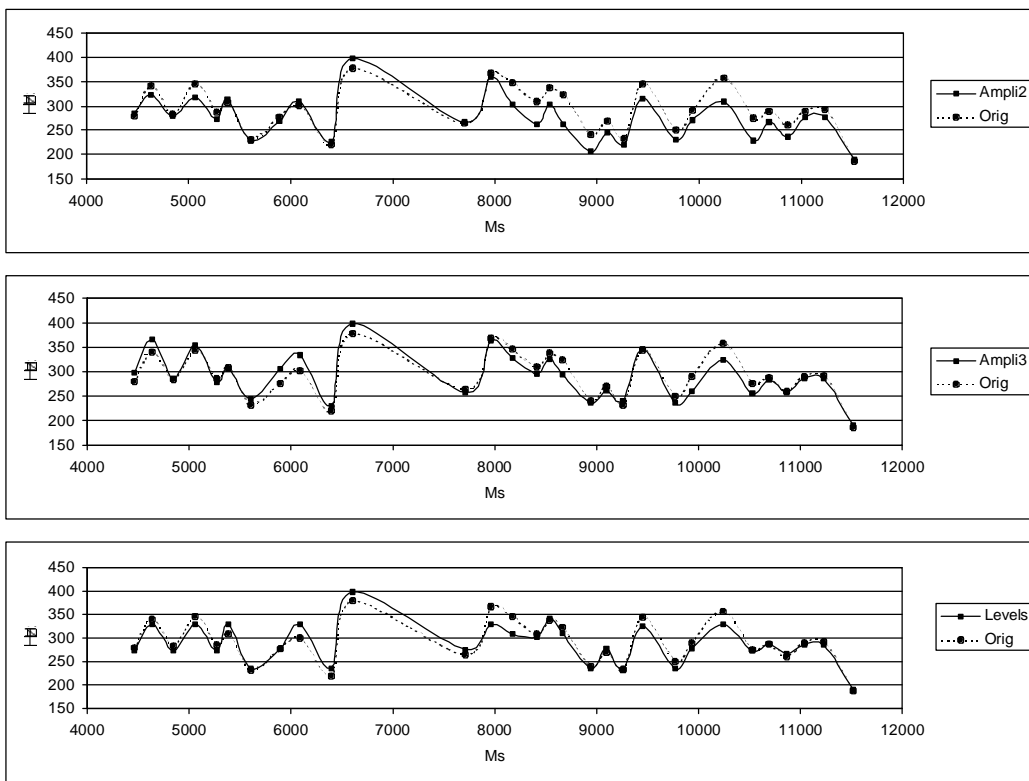


Figure 0.12 (part b). Regeneration with models *Ampli2*, *Ampli3* and *Levels*