Automatic Adaptation of the MoMel F0 Stylisation Algorithm to New Corpora

S. Mouline, † O. Boëffard, P.C. Bagshaw

France Télécom R&D, France
{salma.mouline,paul.bagshaw}@rd.francetelecom.com
†IRISA/Cordial – Université de Rennes 1, France
olivier.boeffard@irisa.fr

Abstract
The paper investigates the adaptability of the MoMel (Modelling of Melody) Algorithm [1] to new corpora. A detailed overview of the MoMel algorithm and its parameters are presented. The generality of the default parameter values to new corpora is studied empirically. Two of the parameters, related to window durations, are discovered to be highly corpus dependent. The paper presents a significant reduction in the modelling error (r.m.s. from 8.45 to 6.19 Hz) by automatically adapting these two parameters to the new corpora.

1. Introduction
Speech technologies such as text-to-speech synthesis cannot circumvent the use of prosodic models. Prosody covers information that can be perceived from speech and concerns essentially supra-segmental factors like phone duration, energy, and melody. One of the most prominent features is certainly the melodic information conveyed by the fundamental frequency, \( F_0 \). An \( F_0 \) contour (which corresponds to the time evolution of the vocal folds frequency) is closely related to the linguistic structure of an utterance, but is also influenced by the speaker’s own characteristics: morphology, linguistic environment, phonological strategy, etc.

Whatever theory of intonation is chosen, an \( F_0 \) stylisation step is always needed. This stylisation should be carried out using models that have a minimum complexity (least possible number of parameters) and obtain the maximum efficiency (minimum error). This can be more or less interrelated with phonological assumptions, i.e. the Tilt [2] and Fujisaki [3] models need phonologic features to approximate \( F_0 \) contours.

This article focuses on the model proposed by Hirst & Espesser [1] called TINST/MoMel. Given a raw sequence of \( F_0 \) values, MoMel finds breaks on the curve. The \( F_0 \) contour is approximated by a quadratic spline between two consecutive ruptures. The working assumptions of the MoMel algorithm are reflected by certain meta-parameters like sliding window durations or minimal and maximal \( F_0 \) values. The aim of this work is to verify the generic properties of the meta-parameter values proposed in [1]. An experimental methodology is proposed herein to compare optimal meta-parameter values estimated on validation corpora with those proposed by the standard MoMel algorithm.

The rest of the paper is organised as follows. Section 2 provides an overview of the MoMel algorithm as detailed in [1]. Section 3 presents the methodology employed to adapt MoMel to new corpora. In Section 4, a short conclusion and an opening of possible extensions to the work are given.

2. MOMEL model review
2.1. Algorithm summary
MoMel (MOdelling of MElogy) is an algorithm proposed by Hirst & Espesser [1] to style an \( F_0 \) contour. The algorithm models the fundamental frequency as a sequence of quadratic splines. Let \( F_0[i] \) be the \( i \)th value of the fundamental frequency in the raw sequence, \( 1 \leq i \leq N \). MoMel considers a fundamental frequency contour as the superposition of a micro-prosodic and macro-prosodic profile. The macro-prosodic profile is ideally smooth and continuous and can be modelled by a sequence of target points connected by an interpolation function based on quadratic splines. The residual micro-prosodic profile is disposed.

The algorithm extracts target points from the \( F_0 \) contour. These target points effectively segment the contour into a sequence of quadratic splines.

The algorithm has four stages:

1. Eliminate the irregularities due to errors in the determination of the \( F_0 \) typically observed at the transitions between voiced and unvoiced parts of speech. All changes in frequency higher than a given threshold (5%) compared to their neighbours are removed:

   \[
   \frac{F_{0}[i]}{F_{0}[i-1]} > 0.05 \text{ or } \frac{F_{0}[i]}{F_{0}[i+1]} > 0.05 \text{ then remove } F_{0}[i].
   \]

2. Estimate the candidate target points with a quadratic regression method on a moving analysis window. This second stage proceeds in three operations on each instant index \( i \):

   (a) All the values not between \( H_{\text{sup}} \) – the maximum authorised value for \( F_0[i] \), \( 1 \leq i \leq N \) – and \( H_{\text{inf}} \) – the minimum authorised value for \( F_0[i] \), \( 1 \leq i \leq N \), including pauses and unvoiced regions of speech – are removed within
an analysis window of duration $A$ centred on instant index $i$.

(b) The remaining values are used to calculate a quadratic regression. Observed values of fundamental frequency $F_0[i]$ are compared with values $\hat{F}_0[i]$ estimated from the quadratic regression. The value $\hat{F}_0[i]$ is removed from the contour if the measure $\frac{|F_0[i] - \hat{F}_0[i]|}{\hat{F}_0[i]} > D$. This operation is reiterated until no new values are removed.

3. Select a partition of these candidate target points. The partitioning of the candidate targets is done on a reduction window of duration $R$ centred on each instant index $i$. A boundary is inserted at the instant $i$ if a distance between the two halves of the window of targets is beyond a dynamic threshold. This threshold corresponds to a moving average from the two halves windows, see [1] for more details. The algorithm tries to find a rupture on the statistics of the $t$ and $h$ values considering only the first statistical moments.

4. Reduce the set of candidate targets by merging or eliminating very close points. Within each partition, targets whose normalised distance (see [1]) are greater than one standard deviation from the mean values for the corresponding segment are eliminated. The empirical average of the remaining candidates is calculated and retained as being the final target of the segment.

Note that this model is based on the adjustment of 5 meta-parameters: the lower limit, $Hz\text{inf}$, and upper limit, $Hz\text{sup}$, for authorised values of $F_0$, the analysis window, $A$, the regression error threshold, $D$, and the reduction window, $R$.

2.2. Default estimation of the parameters

The default settings provided by the MoMel program fix $Hz\text{inf}$ at 50 Hz and $Hz\text{sup}$ at 600 Hz. The three other parameters, the analysis window duration, $A$, the regression error threshold, $D$, and the reduction window duration, $R$, are estimated from a small corpus VNV [5], which consists of two sentences containing all French occlusives and fricatives pronounced by five men and five women.

These parameters are fixed using subjective and objective evaluations. For the subjective method, visual and auditory comparisons are used to determine the parameter values. For the objective method, a distance is calculated between the original $F_0$ contour and the stylised $\hat{F}_0$. Thus the default values selected are: $A = 300$ ms; $D = 0.05$; $R = 200$ ms.

These parameters are then used for the stylisation of other French corpora [1, 4] and the percentage of errors is 5% higher than that for the corpus VNV.

This set of parameters are also used for the stylisation of another corpus [5] but, each time, it is necessary to include a phase of manual correction to improve the quality of the stylisation.

3. Adaptation of the parameters to new corpora

3.1. The corpora

In order to evaluate the variation of the quality of the stylisation of the $F_0$ contours as a function of the meta-parameters, experiments are undertaken on learning and test corpora. These corpora are extracted from a French corpus JPP (a native adult male speaker) made up of 7350 expressively read sentences. The synchronous $F_0$ values are calculated from temporal pitch marks estimated with a standard auto-correlation method followed by a tracking algorithm.

These contours are then de-step filtered, to correct the possible detection errors (this renders stage 1 of MoMel redundant), and sampled at 10 ms. Thus the durations of the analysis and reduction windows are transposed to a number of $F_0$ sample points. So, the default window durations, $A$ and $R$, became lengths $Lfen1=30$ and $Lfen2=20$ respectively.

10 learning corpora, each containing 100 sentences, are made up by randomly pulling out sentences in corpus JPP (no sentences are repeated). Each one of these corpora is used to evaluate the optimal couple $(Lfen1, Lfen2)$ which minimises the quadratic error between the original $F_0$ and that modelled and synthesised by the MoMel algorithm.

20 test corpora are made up in the same manner as the learning ones. These test corpora are used to test the optimal couples estimated by the experiments described below.

3.2. Experiments

3.2.1. Finding optimal meta-parameters

Informal preliminary experiments we have conducted show that the minimum and maximum values for $F_0$ and the regression error threshold $D$ are found to be robust, they are fixed to: $Hz\text{inf}=0$ Hz, $Hz\text{sup}=300$ Hz and $D = 0.05$.

Experiments described in this section aim to evaluate the default MoMel meta-parameters $(Lfen1=30, Lfen2=20)$ and to find the optimal couple of parameters which increases the quality of the stylisation.

The following procedure is undertaken on each of the 10 learning corpora. Given one learning corpus, for each of the 100 sentences, the 2-dimensional space of the parameters $Lfen1$ and $Lfen2$ is swept through the intervals $[5, 40]^2$.

For each couple $(Lfen1, Lfen2)$, the MoMel algorithm is applied on the values of $F_0$ sampled at 10 ms intervals. This stage provides the target points that cut the $F_0$ contour into sequential quadratic splines.
The target points are then used to synthesise the $F_0$ contour. This step generates the frequency values at the same rate (one value every 10 ms). The estimated $F_0$ values are used to calculate an r.m.s. error between the original frequency and the estimated one on the voiced parts of speech:

$$e = \sqrt{\| F_0 - \hat{F}_0 \|^2}.$$

Figure 1: Histogram of the minimal error for both default MoMel meta-parameters and corpus L1 adaptation.

Figure 1 represents the histogram of the default error (dotted line) obtained when the MoMel algorithm is applied with its default parameter values: $(Lfen1=30, Lfen2=20)$ and the minimal error (solid line) estimated on the learning corpus $L1$. The histogram of the minimal error shows the minimal values of error recorded for each sentence while varying the couple of parameters $(Lfen1, Lfen2)$ within the domain $[5,40]^2$.

For each couple $(Lfen1, Lfen2)$, the mean of the errors obtained from all sentences in the 10 corpora are calculated. Figure 2 shows the mean of errors for each couple calculated on the first learning corpus $L1$. It also shows points for the minimal error, which corresponds to the couple $(10,10)$, and for the error when using the default couple $(30,20)$.

These learning experiments demonstrate that the default couple of meta-parameters is not optimal for the corpus JPP. The mean of the r.m.s. error can be reduced from 8.32 to 5.56 by setting $Lfen1$ to 10 and $Lfen2$ to 10.

Note that this optimal couple of meta-parameters has duration windows smaller than the default couple. This is due to the nature of the corpus JPP. As the corpus is composed by expressively read sentences, the $F_0$ contours more turbulent than those of the corpus VNV used to set the default meta-parameters. Thus, the analysis and reduction windows must be shorter so as to model the multitude of peaks and valleys and to take into consideration every significant detail of the $F_0$ contour.

3.2.2. Testing the optimal meta-parameters

The aim of the test experiments is to validate the choice of the optimal couple of the parameters $(Lfen1, Lfen2)$ resulting from the previous learning experiments. These experiments are done on 20 test corpora, each containing 100 sentences.

For each sentence of each corpus, the MoMel algorithm is applied using the optimised values for the couple $(Lfen1, Lfen2)$ as estimated above. The target points are used to regenerate the $F_0$ contour. Then a quadratic error is calculated between the original values and that regenerated by MoMel.

According to the results of the learning experiments for the first learning corpus $L1$, the couples $(10,10)$ reduce the error between the original $F_0$ and that generated by the MoMel stylisation.

Figure 3 shows the histogram of errors obtained for the 100 sentences of each of the 20 test corpora. The solid line histogram is obtained for the couple $(10,10)$, the dotted line corresponds to the default MoMel couple $(30,20)$.

The mean error for the couple $(10,10)$ is 6.19 with a 95% interval of $[5.90, 6.48]$. For the default MoMel couple, the mean error obtained is 8.45 with an interval of $[8.25, 8.64]$.

3.2.3. Variability of the optimum

This section aims to illustrate the variability of the optimal meta-parameters over different learning corpora. The same methodology as exposed in the previous two sections is applied to 10 different learning corpora, the 20 test corpora do not change. For each learning corpus, Figure 4 shows the mean of the r.m.s. error with a 95% confidence interval calculated over the 20 test corpora (the confidence intervals are estimated with a t-student distribution due to the small number of test corpora). The figure also shows the error point corresponding to the default MoMel meta-parameters values.
Note that the differences between all the learning corpora are clearly not significant. The meta-parameter values estimated from the learning corpora outperform the default MoMel values.

A reduction in error (increased efficiency) would not be satisfactory if the enhancement systematically implied more targets (increased complexity). Figure 5 illustrates the influence of the meta-parameters over the complexity of the spline interpolation relating the r.m.s. error to the number of MoMel targets. First, note the randomness of this relation for low error values. Second, there is no clear mapping between the error values and the number of targets for an error value belonging to [5, 10] Hz.

4. Conclusion

This article has presented a methodology to automatically adapt the default MoMel meta-parameter values to a new speaker. The improvement in the modelling error is always significant given a 95% confidence interval (r.m.s. from 8.32 to 5.56 Hz). Despite this optimisation, the observed errors remain perceptually noticeable (JND of approximately 3-4 Hz) and future work needs to address this problem.

5. References


