ModProso:
A Praat-Based Tool for F0 Prediction and Modification

Juan María Garrido[1]

[1] Department of Translation and Language Sciences, Pompeu Fabra University,
Roc Boronat 138, 08018 Barcelona, Spain
juanmaria.garrido@upf.edu

Abstract

In this paper we describe ModProso, a Praat-based tool for prediction and modification of F0 contours in natural utterances. A general overview of the tool is given, and a brief description of the several steps carried out in the F0 contour generation are provided.
Index Terms: F0 contours, Analysis-by-synthesis, Speech Synthesis

1. Introduction

This paper presents ModProso, a Praat-based tool [1] for the perceptual evaluation of 'synthetic' F0 contours predicted from a chain of symbolic labels. It works in a similar way to other existing tools for F0 manipulation and prediction, as ProZed [2], in the sense that it replaces the original F0 contour of a natural utterance by the F0 contour predicted from the intonational labels given as input, but it accepts a different inventory of intonational labels (the ones predicted by the intonational model described in [3,4,5]). It was developed as a research tool to perceptually evaluate the output of the automatic F0 stylisation, annotation and modelling tool described in [5], but it has also been used to generate synthetic stimuli with modified F0 contours for several purposes.

ModProso was originally designed for its use with speech utterances in Spanish and Catalan, but current research in being carried out to adapt it to Brazilian Portuguese and Mandarin Chinese. Adaptation to other languages could be also done with a minimum effort.

2. Background

The tool assumes the model for intonation description proposed in [3,4]. This model conceives F0 contours as the result of the superposition of two types of F0 patterns, as shown in Figure 1

- **Local**: typical F0 shapes occurring at Stress Group (SG) level.
- **Global**: global evolution of an F0 contour along an Intonation Group (IG).

F0 contours can be viewed then as the sum of three types of local patterns, **initial**, **middle** and **final**, depending on its position within the IG, which are superimposed to a global pattern determining its relative height within the speaker F0 range.

Local patterns are modelled as sets of F0 turning points anchored to specific parts of the syllables that make up SG. Each pattern is identified with a label which includes information about:

- the level of the F0 points that make up the pattern: P (Peak), P+ (extra high peak), V (Valley) and V- (extra low valley), depending on the relative height of each F0 point within the F0 range of its container IG;
- the syllable which contains the point: 0 (the stressed one), 1 (one after the stressed one), -1 (one before the stressed one), etc.
- the position of the point within the syllable: I ('initial', close to the beginning of the syllable nucleus), M ('middle', close to the centre of the nucleus), and F ('final', close to the end of the nucleus).

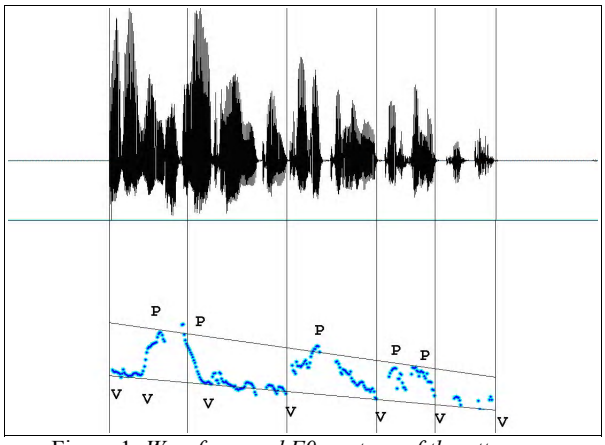

Figure 1: *Waveform and F0 contour of the utterance "Aragón se ha reencontrado como motor del equipo", uttered by a Spanish female speaker. Vertical solid lines represent SG boundaries.*

So for example, a pattern labelled as VI0_PM0_PI1, as the one shown in figure 2, is made up of three F0 inflection points: a V point located at the beginning of the stressed syllable of the container IG (I0); a P point in the middle of the stressed syllable (M0); and a P point at the beginning of the syllable after the stressed one (I1).

Global patterns are modelled as reference lines predicting F0 values as a function of time along the IG, as can be observed in figure 1. The model distinguishes several types of pattern lines (**initial**, **middle** and **final**), according to the position of the IG within its container sentence.

Both local and global patterns for a given utterance can be obtained automatically using MelAn, the modelling tool described in [5]. The tool stores the full listing of local patterns detected in the input utterance within a '.contour' file as the one shown in table 1.

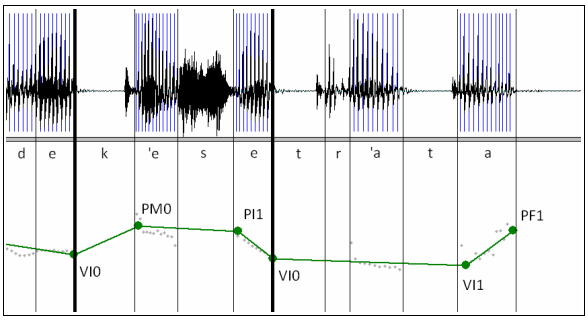

Figure 2: Notation example of two F0 patterns of the utterance '¿Quiere alguien explicarme de qué se trata?', uttered by a female speaker. Vertical solid lines represent SG boundaries.

VI-1_PI0_VI1_INICIAL_4.pattern
VI0_PM0_PM1_INTERIOR_4.pattern
VI0_VF0_FINAL_1.pattern
VF-3_PM-1_VF0_INICIAL_4.pattern
0_INTERIOR_3.pattern
VI0_PM0_INTERIOR_3.pattern
PI0_VM0_PI1_VM1_PI1_VF1_PI2_INTERIOR_3.pattern
VI0_VF1_INTERIOR_2.pattern
PM0_P+I1_PI2_INTERIOR_4.pattern
PI1_INTERIOR_3.pattern
VI2_INTERIOR_3.pattern
VF0_PI2_INTERIOR_3.pattern
VM0_PI1_INTERIOR_2.pattern
VF0_PI1_VI2_PF3_INTERIOR_4.pattern
PI0_VM0_VF0_FINAL_ENUNCIADO_1.pattern

Table 1. *Example of 'contour' file containing the list of F0 patterns for the Brazilian Portuguese utterance 'Depois de tanto caminhar, amanheceu dia na luz da manhã descobriram trinta ou quarenta moinhos de vento que há no Campo de Montiel', spoken by a female speaker.*

Reference lines for global patterns are calculated as two regression lines approaching the P and V points respectively of any IG found in the input utterance. The result is stored in two separate files ('regression_P' and 'regression_V'), which contain the initial F0 of the calculated line, and its slope, in the format shown in table 2.

"x"	
"(Intercept)"	257.848602365733
"Tiempo"	-6.86524557236724

Table 2. *Example of 'regression_P' file containing the initial F0 value and slope for P regression line of the Brazilian Portuguese utterance Depois de tanto caminhar, amanheceu dia na luz da manhã descobriram trinta ou quarenta moinhos de vento que há no Campo de Montiel', spoken by a female speaker.*

These three generated files ('contour', 'regression_P' and 'regression_V') can be used directly as input for ModProso to make 'analysis-by-synthesis' modification of F0 contours.

3. Description of the tool

3.1. General Overview

ModProso performs basically two tasks:

1. the prediction of a chain of F0 target values, in the form of a Praat-style stylised contour, using as input a list of local pattern labels contained in a 'contour' file, and two P and V reference lines, stored as regression lines in 'regression_P' and 'regression_V' files;
2. the substitution of the original F0 contour by the predicted one in the speech utterance provided as input.

To perform these two tasks, ModProso needs as input:

1. a wav file containing the utterance to be manipulated;
2. a Textgrid file containing the orthographic and phonetic transcription of the provided utterance, and its prosodic segmentation into syllables, SG, IG and breath groups (BG), as shown in Figure 3;
3. a 'contour' text file containing the list of pattern labels;
4. a couple of 'regression_P' and 'regression_V' files containing the values for the global F0 reference lines.

All three 'contour', 'regression_P' and 'regression_V' files required as input can be both files obtained from the automatic analysis of a given utterance using MelAn, or 'theoretical' files artificially built for research purposes.

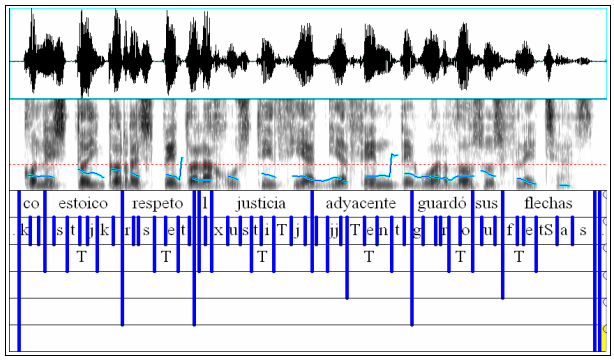

Figure 3: *Speech waveform and TextGrid file for the Spanish utterance 'Con estoico respeto a la justicia adyacente guardó sus flechas', uttered by a male speaker. It contains the necessary tiers to be used as input by ModProso: orthographic transcription (tier 1), phonetic transcription (tier 2), syllable segmentation (tier 3), SG segmentation (tier 4), IG segmentation (tier 5) and BG segmentation (tier 6).*

Figure 3 presents a workflow diagram representing the processing steps frm an input wav file to a new audio file containing the original utterance modified with the predicted F0 contour.

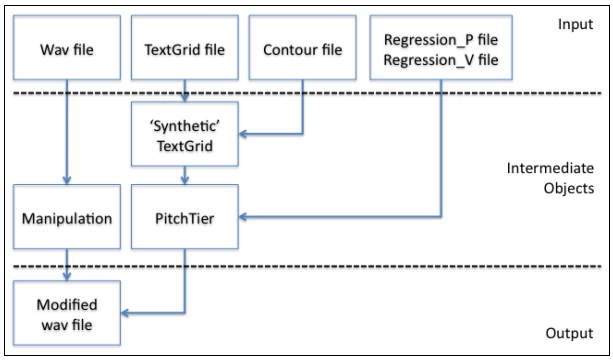

Figure 4: *Workflow diagram showing the processing steps in the generation of a wav file with a predicted F0 contour using ModProso.*

At the end of the process, an edition window showing the speech signal and the obtained stylised contour, as the one shown in figure 5, is displayed. Using this window, the user can obtain a synthesised version of input utterance using Overlap-Add or LPC techniques, which can be played directly or stored in an output wav file.

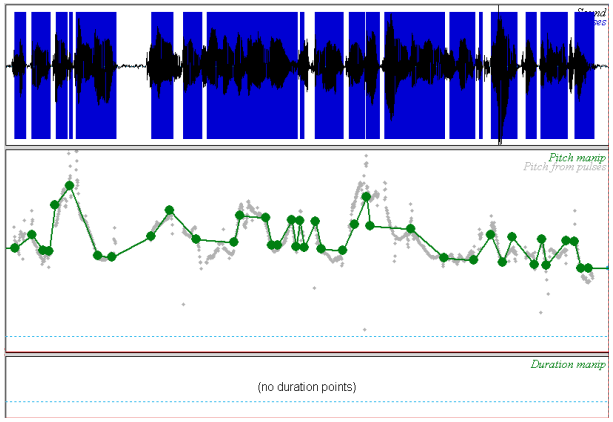

Figure 5: *Praat edition window showing the speech signal and the predicted F0 stylised contour for the Brazilian Portuguese utterance 'Depois de tanto caminhar, amanheceu dia na luz da manhã descobriram trinta ou quarenta moinhos de vento que há no Campo de Montiel', spoken by a female speaker.*

3.2. Prediction of the F0 chain

The generation of a 'synthetic' F0 contour is carried out in two steps:

- **Label alignment**: the labels contained in the 'contour' file are anchored to the predicted places in syllables within its corresponding SG of the input utterance.
- **F0 calculation**: F0 values for each inflection point predicted by the labels are calculated using the P and V regression lines.

In the label alignment phase, a new point tier is added to the input TextGrid showing the alignment of the labels with the signal, as shown in figure 6, to generate an intermediate TextGrid file ('*TextGrid_generado*'), to be used in the F0 calculation process.

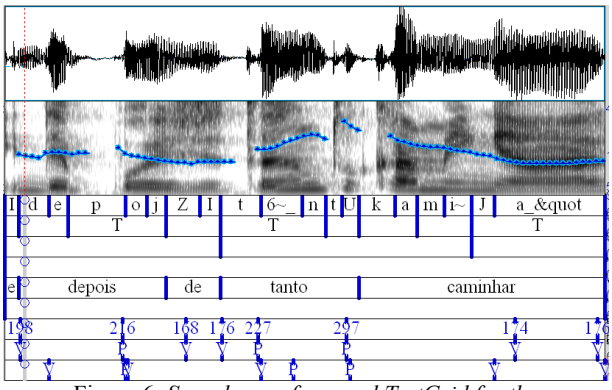

Figure 6: *Speech waveform and TextGrid for the Brazilian Portuguese utterance 'e depois de tanto caminhar', uttered by a female speaker. Last tier shows the predicted alignement of the input labels; they can be compared with the ones obtained with MelAn, appearing in the previous tier*

In this second phase, F0 values for each P and V value are calculated using their predicted time alignment value stored in the intermediate 'TextGrid_generado' file and the regression lines provided as input. After this process, the obtained chain of F0 values is converted into a PitchTier object and then stored in a second intermediate file ('PitchTier_sintetico'), that will be used in the final F0 contour substitution process.

3.3. F0 contour substitution

Finally, the contour substitution process, which is carried out in two steps, takes advantage of the F0 manipulation facilities available in Praat:

1. A 'Manipulation' Praat object is created from the input wav file.
2. The original F0 contour in the obtained 'Manipulation' object is replaced by the PitchTier loaded from the intermediate 'PitchTier_sintetico' file. This modified 'Manipulation' object is the one which is presented to the user in the final edition window.

4. Applications

ModProso has shown to be useful to perceptually evaluate the symbolic representation of F0 contours automatically obtained with MelAn. The results of the experiments carried out with Spanish and Catalan speech corpora, presented in [5], showed that listeners evaluated the synthetised F0 contours as reasonably similar to the original ones, both in Spanish (mean rate 4.05 over a maximum of 5) and Catalan (mean rate 3.93). A similar experiment is being designed to carry out the same evaluation for Brazilian Portuguese.

ModProso has also been used for the generation of manipulated F0 stimuli in other perception experiments, such as the one described in [6], in which the perceptual interpretation of some final F0 patterns used in emotional speech in Spanish was evaluated.

5. References

[1] Boersma, P. and Weenink, W., Praat: doing phonetics by computer [Computer program] http://www.praat.org/, 2012.

[2] Hirst, D., ProZed: A speech prosody analysis-by-synthesis tool for linguists, Speech Prosody 2012. Online:

http://www.speechprosody2012.org/uploadfiles/file/sp2012_submission_70.pdf, accessed on 24 Apr 2013.

[3] Garrido, J. M., Modelling Spanish Intonation for Text-to-Speech Applications, Ph. D Thesis, Universitat Autònoma de Barcelona, 1996. Online: http://www.tdx.cat/handle/10803/4885;jsessionid=376A9A0BED1D5E6DED7CDFD3880316F3.tdx1, accessed on 24 Apr 2013.

[4] Garrido, J. M., "La estructura de las curvas melódicas del español: propuesta de modelización", Lingüística Española Actual, XXIII/2, 173-209, 2001.

[5] Garrido, J. M., "A Tool for Automatic F0 Stylisation, Annotation and Modelling of Large Corpora", Speech Prosody 2010: 100041. Online: http://speechprosody2010.illinois.edu/papers/100041.pdf, accessed on 24 Apr 2013.

[6] Garrido, J. M., Laplaza, Y. and Marquina, M., "On the use of melodic patterns as prosodic correlates of emotion in Spanish", Speech Prosody 2012, Shanghai, 2012. Online: http://www.speechprosody2012.org/uploadfiles/file/sp2012_submission_57.pdf, accessed on 24 Apr 2013.