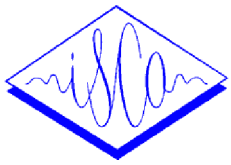
ISCA Archive
http://www.isca-speech.org/archive



Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

A Tool for Automatic F0 Stylisation, Annotation and Modelling of Large Corpora

Juan María Garrido Almiñana

Departament de Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra, Barcelona, Spain

juanmaria.garrido@upf.edu

Abstract

In this paper, an automatic tool for the modelling of F0 contours is presented. The tool is based on Praat, and allows stylisation and annotation of contours, and the definition of global and local patterns, according to the modelling framework described in [1, 2]. The different phases of the process and its application to two corpora of neutral speech are described, and the results of a perception test designed to evaluate to what extent the modelling procedure correctly captures the relevant movements of the F0 contours are presented.

Index Terms: fundamental frequency, automatic modelling, prosody, corpus annotation

Introduction

The goal of this work is to describe a tool for the automatic stylisation, annotation and modelling of F0 contours. It has been conceived for the automatic processing and analysis of large corpora, and it is intended to be speaker and language independent.

Stylisation, annotation and modelling are common tasks in the phonetic analysis of F0 contours: stylisation involves the reduction of F0 contours to a small set of relevant F0 points; in annotation, F0 events are linked to (usually theory-dependent) predefined symbols; finally, modelling involves the definition of a set of relevant F0 patterns from the analysis of a (usually large) set of F0 contours. These processes are usual in both theoretical intonation analysis and F0 processing for speech technologies (specially TTS).

The underlying framework

The F0 modelling framework used to develop this tool is an evolution of the one proposed in [1, 2]. This modelling approach, strongly inspired in the IPO methodology [3], has as final goal the definition of F0 patterns from the phonetic analysis of F0 contours.

In this model, F0 contours are considered to be the result of the combination of two different kinds of patterns:

- global, representing the evolution of the contours at Intonation Group (IG) level;
- local, predicting the evolution of F0 at Stress Group (SG) level; SGs are formed by one stressed syllable and all the following unstressed syllables before the next stressed one (or the end of the IG).

To define these patterns, the model applies a methodology for the representation and analysis of F0 contours which includes the following steps:

- stylisation of the F0 contours;
- annotation of the stylised contours;
- definition of the local and global patterns (modelling) from the annotated contours.

The representation of F0 contours

F0 contours are represented in this model as series of relevant inflection points, obtained after a stylisation procedure. Each inflection point is assigned during the annotation process to the P (Peak) or V (Valley) level, regarding to its relative F0 height within its container IG. Figure 1 shows an example of this kind of representation. F0 labels have been added manually to the figure at the relevant inflection points.

Two extra labels, P+ and V-, are used to mark inflection points showing F0 values clearly higher or lower than the P and V mean levels, respectively. Figure 2 gives an example of typical F0 contour of an interrogative utterance, showing a P+ point at the end of the final rising movement.

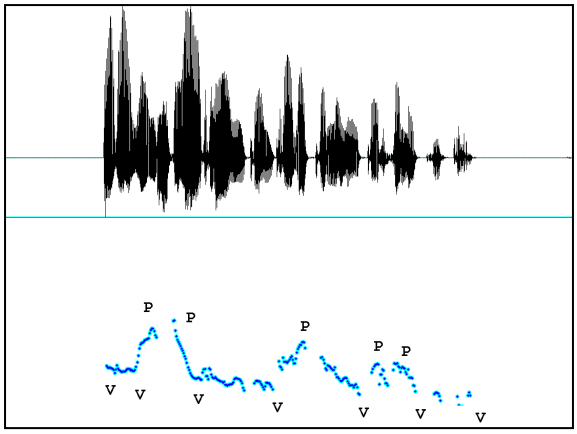


Figure 1: Waveform and F0 contour of the utterance "*Aragón se ha reencontrado con el motor del equipo*", uttered by a Spanish female speaker.

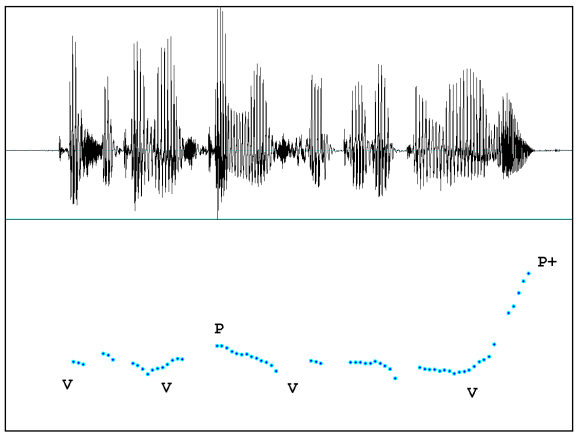


Figure 2: Waveform and F0 contour of the utterance "*¿Que si conozco a mi futuro cuñado?*", uttered by a Spanish male speaker.

Global patterns

Global patterns are represented in the model as 'reference lines' showing the evolution of the P and V F0 levels along the IG. For each IG, then, two reference lines are considered, one for the P and one for the V level. These patterns are speaker-dependent, and model the F0 height and range of each speaker. Figure 3 illustrates this concept. F0 labels and reference lines have been added manually to the figure.

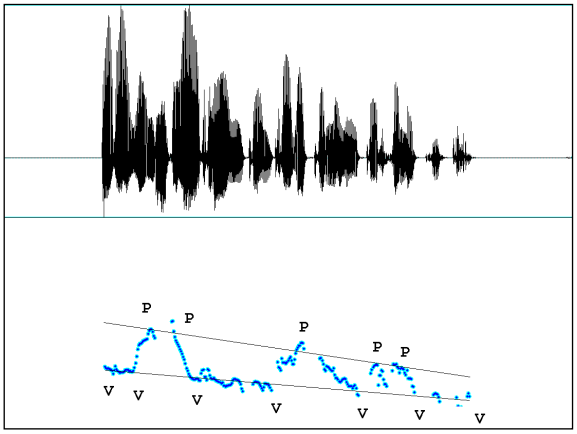


Figure 3: Waveform and F0 contour of the utterance "*Aragón se ha reencontrado como motor del equipo*", uttered by a Spanish female speaker.

Local patterns

Local patterns are defined as recurrent series of labels anchored at specific places of the syllables that make up SGs. The position of each point is defined with respect to the nucleus of its container syllable. Three different positions are considered: 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 4 shows an example of such kind of patterns. Vertical lines represent the boundaries between SGs, delimiting the different local F0 patterns of the contour.

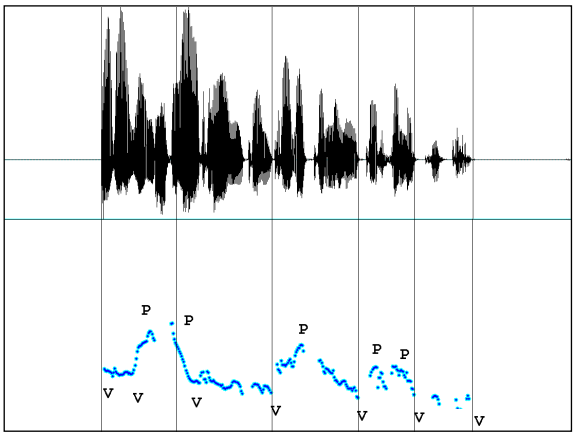


Figure 4: Waveform and F0 contour of the utterance "*Aragón se ha reencontrado como motor del equipo*", uttered by a Spanish female speaker.

The tool

The current implementation of the tool is a set of Praat [4] and R [5] scripts that perform the tasks of F0 stylisation, labelling and modelling. They can be run on Windows, Mac OS X or Linux.

Input

For each utterance to be processed, the tool expects two files as input, a 'wav' file containing the sound and a Praat 'TextGrid' file containing a set of tiers with the orthographic representation of the input utterance, its corresponding phone segmentation aligned with the wav, and the segmentation in prosodic units. To obtain this prosodic segmentation, a different automatic tool, not described here, has also been developed. Figure 5 illustrates the appearance of these files when edited with Praat. The different tiers of the TextGrid represent, in this order, word, phone, syllable, SG, Intonational Phrase (IP) and IG segmentation.

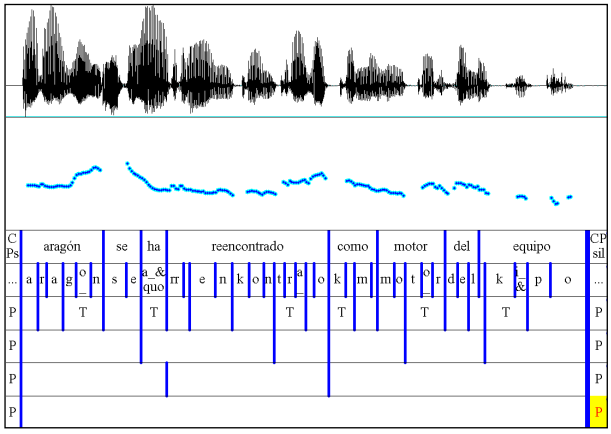


Figure 5: Waveform, F0 contour and TextGrid of the utterance "*Aragón se ha reencontrado como motor del equipo*", uttered by a Spanish female speaker. The Textgrid contains the tiers required as input by the automatic tool.

Processing steps

The processing procedure involves three steps: stylisation, annotation and modelling. Each one is described in more detail in the following subsections.

3.1.1. Stylisation

The stylisation phase detects the relevant F0 inflection points from the original F0 contours. The stylisation procedure included with Praat, tuned properly to obtain a stylised contour perceptually close to the original one, is used for this task. The script also converts the obtained stylised contour to a point tier and adds it to the input Textgrid. Figure 6 illustrates the output of this step. The last tier of the TextGrid shows the output of the stylisation process.

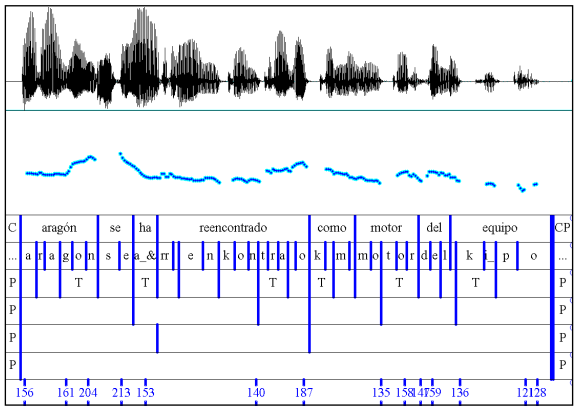


Figure 6: Waveform, F0 contour and TextGrid of the utterance "*Aragón se ha reencontrado como motor del equipo*", uttered by a Spanish female speaker. The last tier of the Textgrid includes the F0 values retained after the stylisation task.

3.1.2. Annotation

The annotation phase involves several steps:

- calculation of regression lines for the F0 values of each IG, using R, to define the boundary between the P and V levels;
- a first labelling loop, to assign P and V labels to each inflection point of the stylised contour;
- calculation of separate regression lines for the P and V points of each IG, again with R;
- a second labelling loop, to assign P+ and V- labels to those points whose F0 value is too far away from the P or V regression line, respectively;
- a final loop of deletion of redundant labels.

The output of this phase is a chain of P+, P, V and V- labels associated with specific inflection points of the contour (not all inflection points receive a label at the end of the process), which are added to the TextGrid in a new point tier. Figure 7 illustrates the output of the annotation process. Last three tiers of the TextGrid show the output of different annotation steps. Last one is the final output of the annotation phase.

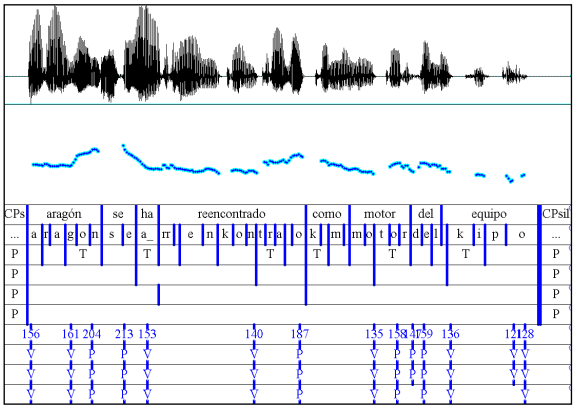


Figure 7: Waveform, F0 contour and TextGrid of the utterance "*Aragón se ha reencontrado como motor del equipo*", uttered by a Spanish female speaker. The last three tiers of the Textgrid show the output of the different steps of the labelling task. The last tier is the one used for the modelling task.

3.1.3. Modelling

According to the modelling approach previously described, two types of patterns are obtained during this phase:

- local patterns, corresponding to series of labels associated to SG;
- global patterns, corresponding to P and V regression lines for each IG of the utterance.

During the local pattern extraction phase, an exhaustive list of the local F0 patterns appearing in the input corpus, as well as their number of appearances, is defined. Also, for each analysed utterance, a 'contour' file is generated, which contains the list of SG patterns detected at its corresponding stylised F0 contour.
As far as global patterns is concerned, individual P and V regression lines are re-calculated (excluding P+ and V- values) for each IG of the analysed corpus. Each regression line is defined by its initial value (in Hz) and its slope (in Hz/sec).

Output

The output of the stylisation and annotation procedures is an enriched TextGrid containing the stylisation and annotation output, as illustrated in figure 7. For modelling, the output is the set of 'contour' files, the regression lines for each IG of the corpus , and a set of summary tables containing the list of local patterns for each type of SG with their corresponding frequency in the corpus, and the mean regression lines for the different types of IG found in the corpus (table 4). This output can be used for analysis purposes, but it is also conceived for its use as input for an automatic F0 contour generation module.

Perceptual evaluation

As mentioned before, this procedure has been applied to the automatic annotation and modelling of several corpora in Spanish and Catalan, and the tool has shown to be robust and effective.
However, in order to evaluate to what extent the applied procedure is able to capture the perceptually relevant movements of F0 contours, a perceptual evaluation test was designed and carried out on a panel of selected listeners. The goal was also to determine if the tools offers similar results when applied to different speakers and languages (Spanish and Catalan).

Test design and procedure

For this test, a set of 20 utterances per language, 10 uttered by a male speaker and 10 by a female speaker, was selected. Utterances were selected from different corpora, so they were different from male to female speakers. It was also intended during the selection of utterances to have a mix of short and medium-sized utterances, to check the effect of utterance length on the procedure. Long sentences, however, were avoided because they are difficult to evaluate in a perception test.
For each selected utterance, two synthesised versions were created: a first one using the original F0 contour, and a second one from the local and global patterns automatically defined by the tool for those utterances. Figures 8 and 9 give an example of both types of stimuli. To generate these versions, a Praat-based synthesis tool which allows to built stylised F0 contours from 'contour' files and regression lines, and to generate synthesised versions of the utterance using this modelled contour, was used. In order to minimize the differences between the 'original' and 'predicted' versions due to the used synthesis technique, the LPC-based synthesis procedure included in Praat was used instead of the Overlap-Add one.

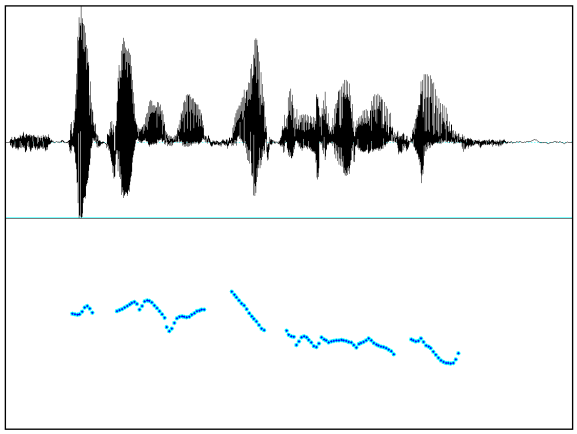


Figure 8: Waveform and F0 contour of the synthesised version of the utterance "*Y cada vez la tendremos más*", uttered by a Spanish female speaker. The F0 contour used to generate this version is the original one.

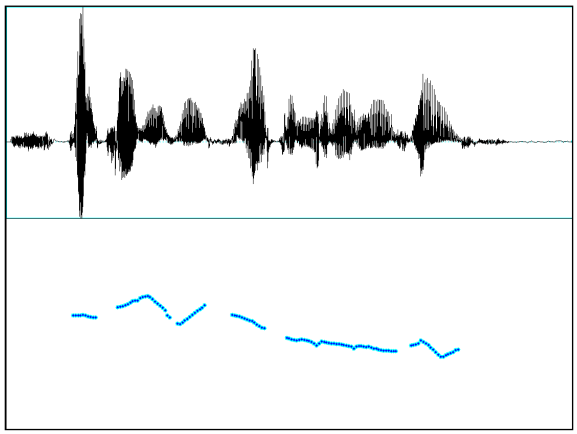


Figure 9: Waveform and F0 contour of the synthesised version of the utterance "*Y cada vez la tendremos más*", uttered by a Spanish female speaker. The F0 contour used to generate this version is the modelled one.

A group of 10 judges (5 experts and 5 non-experts) was asked to listen to the stimuli and to rank the degree of similarity of both intonations in a 1-5 scale. The test was accessible via web, and they were told to use headphones to run it.

Results

The results of the perception tests for Spanish and Catalan presented in tables 1 and 2 show that in general listeners perceived as quite similar the original and the modelled F0 contours presented for each utterance (mean global scores of 4.05 for Spanish and 3.93 for Catalan). There are of course differences among utterances (mean scores per utterance ranging from 4.9 to 2.4 in the case of Spanish, and from 4.8 to 1.6 in the case of Catalan), but very few utterances showed a mean score below 3: sentences 4 and 5 (2 out of 20) in the case of Spanish; sentences 3, 11 and 17 (3 out of 20) in Catalan. Mean results are a bit lower in Catalan than in Spanish, but a closer analysis to the stimuli revealed that the main source of deviations between the original and modelled versions were errors in the automatic labelling of the inflection points (assignment of P labels instead of V, or P instead of P+, for example), not to intonational differences between the two languages.

Discussion, conclusions and future work

The results presented in the previous section lead to state that the stylisation and annotation tool presented in this paper allows to obtain modelled versions of the F0 contours quite similar to the original ones in a fully automatic way. There are of course problems pending to solve (for example, the labelling errors detected during the evaluation task), but the results are encouraging.

This tool is intended to be used in the next future for the automatic prosodic annotation of large corpora, within the framework of the GLISSANDO project. Also, experiments in the automatic generation of F0 contours for synthesis purposes are planned in the next future.

Table 1: Utterance-by-utterance and global mean rates for the Spanish test

Utterance number	Average rating
1	4.8
2	3.9
3	3.1
4	2.8
5	2.4
6	4.6
7	3.9
8	4.2
9	4.5
10	4.4
11	4.5
12	4
13	4.5
14	4.9
15	4.2
16	4.7
17	4
18	3.5
19	3.8
20	4.3
Total	**4.05**

Table 2: Utterance-by-utterance and global mean rates for the Catalan test

Utterance number	Average rating
1	4.2
2	4.1
3	1.6
4	4.5
5	4.8
6	4.7
7	4.7
8	4.1
9	3.7
10	3
11	2.4
12	3.3
13	4.2
14	4.2
15	4.5
16	4.8
17	2.8
18	4.4
19	4.4
20	4.3
Total	**3.935**

Acknowledgements

The author would like to thank Yesika Laplaza, Montserrat Marquina and Junming Xiaoyao for their comments to a previous version of this paper.

References

[1] Garrido, J. M., Modelling Spanish Intonation for Text-to-Speech Applications, Ph. D. Thesis, Departament de Filologia Espanyola, Universitat Autònoma de Barcelona, 1996; http://www.tesisenxarxa.net/TDX-0428108-155145/index_cs.html, accessed on 12 Nov 2008.

[2] Garrido, J. M., La estructura de las curvas melódicas del español: Propuesta de modelización. *Lingüística Española Actual, 23*(2), 2001, 173-210.

[3] 't Hart, J. – Collier, R. – Cohen, A.. *A perceptual study of intonation. An experimental-phonetic approach to speech melody*, Cambridge, Cambridge University Press, 1990.

[4] Boersma, P., Praat, a system for doing phonetics by computer. *Glot International* **5:9/10**, 2001, 341-345.

[5] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2005; http://www.R-project.org, accessed on 12 Nov 2008.