SegProso:
A Praat-Based Tool for the Automatic Detection and Annotation of Prosodic Boundaries in Speech Corpora

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 SegProso, a Praat-based tool for the automatic segmentation in prosodic units of speech corpora. It is made up of a set of Praat scripts that add several tiers, each one containing the segmentation of a different unit, to a previously existing TextGrid file including the phonetic segmentation of the associated wav file. It has been successfully used for the annotation of several corpora in Spanish and Catalan. The paper briefly describes the workflow of each detector, and presents the results of an evaluation of the performance of the tool in an automatic annotation task on two small Spanish and Catalan corpora.

Index Terms: Prosody, Speech Corpora, Automatic Annotation

1. Introduction

The annotation of prosodic boundaries in speech corpora is a time-consuming task if it is performed by manual means, especially in the case of the annotation of large corpora; and in many cases there can be a strong inter-annotator disagreement in the perceptual identification of some units. Automatic annotation is a promising alternative solution for this problem: even if the obtained output is not perfect, it allows to reduce significantly the time devoted by human experts to this task.

In this paper we describe SegProso, a Praat-based tool [1] for the automatic segmentation of speech corpora into prosodic units. It is made up of a set of Praat scripts which add to a previously existing TextGrid file four tiers containing the segmentation into **syllables**, **stress groups** (SG), **intonation groups** (IG) and **breath groups** (BG). The tool uses a rule and knowledge-based approach to perform the boundary detection tasks. It was originally designed for the annotation of speech in Spanish and Catalan, but current research in being carried out to adapt the tool to Brazilian Portuguese, and languages could also be added with minimum or none adaptation of the scripts, if the corresponding phonetic transcription was provided. The tool is available for public download at http://www.upf.edu/pdi/jmgarrido/recerca/projectes/segproso.zip.

In the following pages, an overview of the tool is given, and the different scripts in charge of the detection of the each type of unit are described. Also, the results of an informal evaluation of the performance of the tool for the automatic annotation of a small speech corpus in Spanish and Catalan are provided.

2. Description of the tool

2.1. General Overview

SegProso is made up of a set of four Praat scripts, each one performing a different segmentation task:

- *Syllable boundaries detector*
- *SG boundaries detector*
- *IG boundaries detector*
- *BG boundaries detector*

These scripts can be run sequentially, to perform a full annotation task, or in isolation, to annotate a single level, if the necessary input for each script is provided: a wav file containing the speech signal of the utterance to be annotated, and a Praat TextGrid file containing the necessary tiers to perform the annotation task.

Full annotation using SegProso can be done by running another script which makes sequential calls to each individual detection script, in the necessary order to ensure that each one will find the necessary input information in the TextGrid file: syllable annotation is performed first; then IG annotation; next is SG annotation; and finally BG annotation.

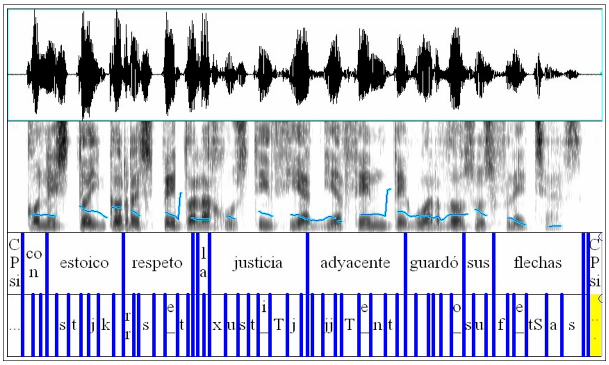

Figure 1: *Speech waveform and TextGrid containing the word segmentation (tier 1) and phonetic segmentation (tier 2) corresponding to the utterance 'con estoico respeto a la justicia adyacente guardó sus flechas', spoken by a male speaker.*

The TextGrid file provided as initial input to SegProso must contain at least two tiers: an interval tier the first one including the orthographic transcription of the utterance word-by-word; and a second interval tier with the phone segmentation in SAMPA format [2]. Figure 1 provides an example of such an input. Pauses must be also marked and annotated with specific label in both tiers.

The tool provides as output the same input TextGrid file enriched with four new tiers, containing the segmentation of the four prosodic units mentioned above. Figure 2 offers an example of the appearance of such a file.

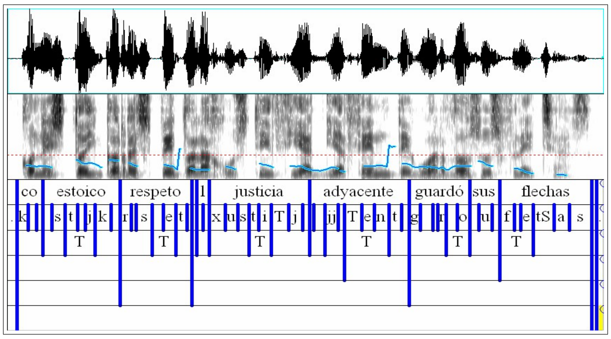

Figure 2: *Speech waveform and TextGrid for the same utterance of Figure 1 containing the output tiers of SegProso: syllables (tier 3), SG (tier 4) IG (tier 5) and BG (tier 6).*

2.2. Syllable boundaries detector

The syllable detection script creates a new tier in the input TextGrid with the syllable boundaries corresponding to the phone chain, in SAMPA transcription, provided as input in the same TextGrid. It also annotates the intervals corresponding to stressed syllables with a specific label.

To perform this task, the script implements a set of linguistic rules which predict the grouping of phones appearing in the input phonetic transcription tier. The general workflow of these rules is as follows:

1) The script first locates word boundaries in the orthographic tier: word boundaries are assumed to be a 'barrier' for phone grouping, so it is carried out separately within each word.
2) The script scans then the input phone chain of each word in search of phone symbols representing syllabic nuclei. The procedure in charge of this task basically checks if the input symbol appears in the implemented list of 'nuclear' phones (initially only Spanish and Catalan vowel symbols, recently enlarged with those of Brazilian Portuguese vowels; only vowels are allowed to be syllabic nuclei in those languages).
3) Once a nucleus has been detected, the script tries to define the boundaries of the corresponding syllable. To do this, it looks further in the phone chain, detects if there are non-nuclear phone combinations before the next nucleus and if so, tries to establish the placement of the final syllabic boundary by applying the corresponding syllabication rules. As already mentioned, a word boundary is considered always to be a syllable boundary as well; no re-syllabication procedures across words are applied.
4) If the nucleus contains a stressed vowel (it as to be transcribed as stressed in the phone tier to be identified), the syllable is labeled as 'stressed' in the corresponding interval of the output tier (label 'T', for 'Tonic').

This approach is different from other existing tools, such as APA [3], in which syllable boundary detection is attempted from the acoustic analysis of the speech signal. In this case, a 'theoretical' grouping of phones in the transcription tier is done based on phonological syllabication rules, not on the detection of acoustic cues for syllable boundaries.

The phone grouping rules implemented in the script are intended to be language-independent, in the sense that they try to represent phonological grouping principles valid at least for several Romance Languages. However, its current implementation is language-dependent, in the way that it uses language-dependent phone inventories in the nuclei detection and syllabication rules. The adaptation of the script to perform syllable segmentation in Brazilian Portuguese involved only the addition of new symbols to the inventory of possible syllable nuclei, with no extra syllabication rules, but probably the adaptation to other languages would not be so direct.

A similar approach has been described recently in the implementation of the syllable annotator of the SPPAS system [4], although in that case syllabication rules seem to be fully language-dependent.

2.3. Intonation group boundaries detector

IG is usually defined as the natural domain of a 'complete' intonation contour. A contour is considered to be complete if it is closed with a final (boundary) F0 pattern, a pause, or both. Other additional phonetic cues, such as declination resets or pre-boundary syllable lengthening, may also indicate the presence of an IG boundary. Some theoretical approaches make a distinction between major (usually ended with a pause) and minor IG (no pause, only boundary pattern at the end). Perceptual identification of IG boundaries is sometimes difficult by non-expert listeners. For this reason, manual segmentation is usually difficult, showing important inter-annotator disagreement.

The script in charge of the identification of IG boundaries in SegProso needs to have in the input TextGrid three tiers containing the word and syllable boundaries, and the phonetic transcription, as well as the corresponding wav file. It tries to detect boundaries at the end of stressed words, which are the candidate places, by looking for two types of F0 cues: the existence of specific F0 boundary patterns, on the one hand, and the existence of declination resets, on the other. The segmentation process is carried out by two sets of rules that look for specific differences between F0 values at specific syllables before and after word boundaries:

- Boundary pattern rules look for specific F0 risings just before word boundaries that could be perceptually interpreted as 'boundary' movements. Basically, these rules compare F0 values in the nucleus of the stressed syllable with F0 values in the last post-stressed syllable (if any), or at the end of the same stressed syllable, if it is the last syllable of the word. If the difference between these two values is beyond a fixed threshold (currently, 5% of the value at the center of the stressed syllable), a boundary is set at the end of the word. Figure 3 shows an example of it.
- F0 reset rules look for F0 jumps between the stressed syllables of two consecutive words. F0 values are taken in the middle of the nuclei of both stressed syllables. If the F0 of the second syllable is found to be significantly higher that the one at the first syllable (currently, at least 5% higher than the value at the center of the first

stressed syllable), an IG boundary is found in the word boundary between the two stressed syllables. Figure 4 shows an example.

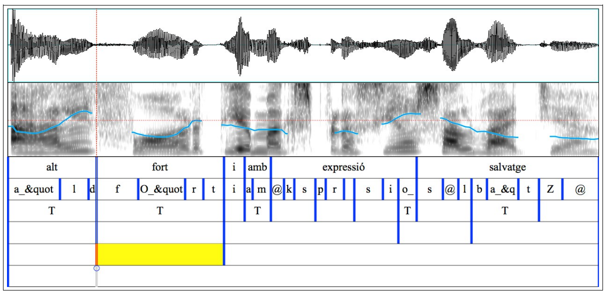

Figure 3: *Speech waveform and prosodic segmentation of the Catalan utterance 'Alt, fort, i amb expressió salvatge', spoken by a female speaker. The selected boundary was inserted by the 'boundary pattern rules' of the IG detection script*

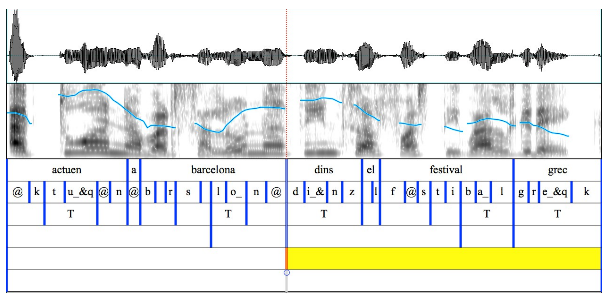

Figure 4: *Speech waveform and prosodic segmentation of the Catalan utterance 'Actuen a Barcelona dins el festival Grec', spoken by a female speaker. The selected boundary was inserted by the 'F0 reset rules' of the IG detection script*

Other segmentation tools, such as APA [3] or ANALOR [5], make use of these F0 cues to detect prosodic breaks similar to IG, sometimes in conjunction with other non-tonal parameters (pauses, energy). However, the detection procedure in SegProso is slightly different, allowing, for example, the identification of F0 resets when no pause is present.

2.4. Stress group boundaries detector

Syllables and IG are two types of prosodic units widely accepted in the Prosodic Phonology literature independently on the theoretical approach. SG, however, is a more theory-dependent prosodic unit, proposed in Garrido [6, 7] and other intonation description frameworks, such as the one by Thorsen [8], for the description of intonation contours. It is defined as a segment of utterance starting at the beginning of a stressed syllable and ending at the beginning of the next stressed syllable, if any, or the end of the container IG. Unstressed syllables appearing at the beginning of an intonation group, before the first stressed one, are considered to be part of the first stress group. Then, for example, the Spanish sentence '*La universidad Pompeu Fabra está en Barcelona*' would be segmented in the following SG:

[La universi**dad** Pom] [**peu**] [**Fa** bra es] [**tá** en Barce] [**lo** na]

The script for the detection of SG needs the syllable and IG segmentation to be already available in the input TextGrid file: it must be run, consequently, after applying the corresponding scripts for the detection of those units. Its workflow is very straightforward: basically, it looks for stressed syllables in the syllable chain (identified with a 'T' in the syllable tier), and places the beginning of each SG at the time marked for the beginning of the syllable in the syllable tier. If the stressed syllable is the first one in the IG, the initial mark of the SG is placed at the beginning of the IG. As far as final marks are concerned, they are placed at the beginning of next stressed syllable, or the end of the IG if the stressed syllable is the last one in the IG.

2.5. Breath group boundaries detector

BG is the last prosodic unit annotated by SegProso. BG are defined as portions of utterances between two silent pauses. They may include one or several IG, or even none (in interrupted utterances, for example).

BG detection script uses the information about the location of the pauses contained in the syllable tier to create a new tier with the segmentation in BG; it only needs then a syllable tier to be present in the input TextGrid file. Its processing workflow is also quite straightforward: the beginning of a new BG is set in the output tier when the end of a pause interval is detected in the syllable tier, and, accordingly, a BG end boundary is detected when the beginning of a new pause interval is found.

3. Evaluation

Two informal evaluation tests, one for Spanish and one for Catalan, were carried out to assess the performance of the tool. The goal of the evaluation was to check to what extent the tool is able to place correctly prosodic unit boundaries in a small automatic annotation task, assuming that the input (phonetic transcription, phone and pause alignment) is correct.

A set of 100 utterances for each language was selected as evaluation corpus. In the case of Spanish, the evaluation corpus was extracted from the Spanish subset of the INTERFACE corpus [9]: 50 utterances spoken by a male and 50 by a female speaker. For Catalan, the utterances were selected from two corpora recorded at Barcelona Media by two different female professional speakers for synthesis purposes [10, 11]. The utterances were rather short, and uttered with a neutral style. TextGrid files obtained automatically using an HMM segmentation tool were available for each fie of the corpus. These utterances were processed using SegProso to obtain their automatic prosodic segmentation. The output was then manually checked, and compared with the automatic version.

The results of the evaluation show an excellent performance of the syllable and BG scripts for both languages: 100% of correct segmentations. This percentage does not include, however, errors in BG segmentation related to some wrong detection of pauses by the automatic segmentation tool. For SG and IG, the obtained rates are lower, although still high. In the case of SG tiers, the percentage of correct boundaries is very similar: 87.32% for Spanish and 86.46% for Catalan. It important to notice that all detected errors were directly related to previous wrong boundary placements in the IG tier. The performance of the script is perfect when the IG boundaries are correctly detected. The script for IG identification is the one which obtained the poorest results: 82.46% of the boundaries inserted by the tool in the Spanish corpus were labelled as correct during the evaluation process, and 77.40% in the case of the Catalan corpus. Some of the wrong boundaries inserted by the tool were moved to another boundary (21 cases, 6.81% of the automatic boundaries, in Spanish; 31 cases, 8.75%, in Catalan) and the rest was deleted (33 cases, 10.71% of the automatic boundaries, in Spanish; 49 cases, 13.84%, in Catalan). During the evaluation process, some boundaries not detected by the tool had to be added manually (22, 7.4% of the correct boundaries, in Spanish; 18, 5.57%, in Catalan).

4. Applications

SegProso has been successfully used for the automatic prosodic annotation of several corpora in Spanish and Catalan, such as Interface, I3Media or Glissando [12]. In all three cases, the obtained segmentation has been used as input for MelAn, the automatic F0 annotation and modeling tool described in [13]. A full inventory of the F0 patterns appearing in those corpora was successfully obtained using this tool. The prosodic segmentation provided by SegProso has also been used for the development of intonation models for TTS [14].

5. Conclusions and future work

SegProso has shown to be a useful tool for fast annotation of prosodic boundaries of large speech corpora in Spanish and Catalan. Although the performance of the IG detection script could be still improved, our experience has revealed that manual revision of the obtained output is much faster than manual annotation.

Future work will focus on the improvement of the IG annotation script: fine tune of the rules is expected to be done using the data of a detailed acoustic analysis of the F0 patterns appearing at those boundaries.

6. References

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

[2] Wells, J. C., "SAMPA computer readable phonetic alphabet", in D. Gibbon, R. Moore, R. Winski, [Eds.], Handbook of Standards and Resources for Spoken Language Systems, Part IV, section B, Mouton de Gruyter, Berlin and New York, 1997.

[3] Cutugno, E., D'Anna, L., Petrillo, M. and Zovato, E., "APA: towards an automatic tool for prosodic analysis", Speech Prosody 2002, 231-234, 2002.

[4] Bigi, B. "SPPAS: a tool for the phonetic segmentation of speech", LREC 2012 Proceedings, 1748-1755, 2012.

[5] Avanzi, M., Lacheret-Dujour, A. and Victorri, B., "Analor: A tool for semi-automatic annotation of french prosodic structure", Speech Prosody 2008, 119–122, 2008.

[6] 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..

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

[8] Thorsen, N., "An acoustical investigation of Danish intonation", ARIPUC, 10, 85-147, 1976.

[9] Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A. and Nogueiras, A., "Interface databases: Design and collection of a multilingual emotional speech database", Proceedings of LREC'02, 2024-2028, 2002.

[10] Garrido, J. M., Bofias, E., Laplaza, Y., Marquina, M., Aylett, M. and Pidcock, Ch., "The Cerevoice speech synthesiser", Actas de las V Jornadas de Tecnología del Habla, 126-129, 2008.

[11] Garrido, J. M., Laplaza, Y., Marquina, M. Pearman, A. Escalada, J. G., Rodríguez, M. A. and Armenta, A., "The I3Media speech database: a trilingual annotated corpus for the analysis and synthesis of emotional speech", LREC 2012 Proceedings, 1197-1202, 2012.

[12] Garrido, J. M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., de-la-Mota, C., González, C., Rustullet, S., Larrea O., Laplaza, Y., Vizcaíno, F., Cabrera, M. and Bonafonte, A., "Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan", Language Resources and Evaluation, DOI 10.1007/s10579-012-9213, 2013. Online: http://link.springer.com/article/10.1007/s10579-012-9213-0, accessed on 24 Apr 2013.

[13] 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.

[14] Garrido, J. M., "GenProso: a parametric prosody prediction module for text-to-speech applications", IberSpeech 2012 Proceedings.