Dialectologia. Special issue, VI (2016), 223-244.
ISSN: 2013-2247

Received 22 February 2016.

Accepted 3 May 2016.

SEA_AP: A SEGMENTATION AND LABELLING TOOL FOR PROSODIC ANALYSIS

Paula LÓPEZ OTERO, Laura DOCÍO FERNÁNDEZ, Carmen GARCÍA MATEO, Marta MARTÍNEZ MAQUIEIRA, Rocío VARELA FERNÁNDEZ & Elisa FERNÁNDEZ REI*

AtlantTIC Research Center for Information and Communication Technologies, Universidade de Vigo / Instituto da Lingua Galega, Universidade de Santiago de Compostela*

plopez@gts.uvigo.es / ldocio@gts.uvigo.es / carmen.garcia@uvigo.es / mmartinez@gts.uvigo.es / rvarela@gts.uvigo.es / elisa.fernandez@usc.es

Abstract

This paper introduces a tool that performs segmentation and labelling of sound chains in phono units, syllables and/or words departing from a sound signal and its corresponding orthographic transcription. In addition, it also integrates acoustic analysis scripts applied to the Praat programme with the aim of reducing the time spent on tasks related to analysis, correction, smoothing and generation of graphics of the melodic curve. The tool is implemented for Galician, Spanish and Brazilian Portuguese. Our goal is to contribute, by means of this application, to automatize some of the tasks of segmentation, labelling and prosodic analysis, since these tasks require a large investment of time and human resources.

Keywords

segmentation, labelling, prosody, alignment, AMPER

SEA_AP: UNA HERRAMIENTA DE SEGMENTACIÓN Y ETIQUETADO PARA EL ANÁLISIS PROSÓDICO

Resumen

En este artículo se presenta una herramienta que realiza la segmentación y el etiquetado de cadenas sonoras en unidades de fono, sílaba y/o palabra partiendo de una señal sonora y de su

©Universitat de Barcelona

correspondiente transcripción ortográfica. Además, integra *scripts* de análisis acústico que se ejecutan sobre el programa Praat con el fin de reducir el tiempo invertido en las tareas de análisis, corrección, suavizado y generación de gráficos de la curva melódica. La herramienta está implementada para gallego, español y portugués de Brasil. Nuestro objetivo es contribuir con esta aplicación a automatizar algunas de las labores de segmentación, etiquetado y análisis prosódico, pues constituyen tareas que requieren una gran inversión de tiempo y de recursos humanos.

Palabras clave
segmentación, etiquetado, prosodia, alineamiento**,** AMPER

1. Introduction

The study of prosody, and intonation in particular, focuses on the description of the evolution of the fundamental frequency, the duration and the intensity linked to the sound chains which convey linguistic meaning. One of the fields of study of the prosody is the analysis of linguistic variation (geographical, social or contextual, mainly). In recent years, the study of the geoprosodic variation has received considerable attention from researchers, so that today we have a significant number of prosodic atlas. In the Romance field, we should highlight the works carried out in the framework of two projects: the *Atlas Multimedia de la Prosodia del Espacio Románico* - AMPER (http://dialecto.u-grenoble3.fr/AMPER/amper.htm) and the *Interactive Atlas of Romance Intonation* - IARI (http://prosodia.upf.edu/iari/index.html). In both projects a great number of European and American universities and research groups collaborate. In both cases, there are numerous results which are available on the different websites (general and specific for each group) as well as in publications of various kinds (Mairano 2011, which also includes a large number of references; Frota & Prieto 2015).

We are particularly interested in the case of AMPER (Contini et al. 2002), as it collects data from all the Iberian varieties (Galician, Asturian-Leonese, Portuguese, Castilian Spanish, Catalan, etc.). Furthermore, in the case of this atlas, it provides a detailed phonetic description of the intonation contours of declarative and interrogative modalities, which has enabled not only an important database based on

©Universitat de Barcelona

data mapping, but it has also facilitated the development of other methodologies of geolinguistic study, such as the dialectometry (Fernández Planas et al. 2011 and 2015, Martínez Calvo & Fernández Rei 2015, Moutinho et al. 2011). Behind all these outcomes, there is intense work comprising recording, codification and analysis at a high cost, both in terms of human resources and time. The tasks of labelling audio files (from which the prosodic data which allows us to perform, amongst other tasks, data analysis or mapping, is obtained) are only partially automatized, making it a very strenuous process, both in developing routines for Matlab, by Antonio Romano (http://www.lfsag.unito.it/amper/fox.html) or by the group AMPER Asturias (http://www.unioviedo.es/labofone/), as in the preparation of scripts in Praat (Boersma & Weenik 2013), in which important contributions have been made by Albert Rilliard (https://perso.limsi.fr/rilliard/InterfaceAMPER.html) and the AMPER group of the Universitat de Barcelona (Martínez Celdrán & Fernández Planas 2003-2015).

This project began circa 2000, so that during almost fifteen years, prosodic dialectal data has been gathered and analysed. Each group has tried to solve the problems that have arisen and there have been many advances from which all of the groups have benefited. Nevertheless, there has also been some fragmentation and dispersal of the resources and tools developed. The need therefore arises to unify these resources in packages by taking the advantages jointly offered.

With regard to the tools for the alignment of audio and text with subsequent labelling of results, various studies have appeared, such as EasyAlign (Goldman 2011), aimed at the alignment of audio and text and with support for Spanish, English, French, Brazilian Portuguese and Taiwanese and which, in addition, is run using Praat. There is also SegProso (Garrido 2013), which is a Praat-based tool for the automatic segmentation of speech corpora into prosodic units for Spanish and Catalan that need initial TextGrid files which must contain at least two tiers, the first of which with the word segmentation and a second tier including a phonetic transcription in SAMPA format. This tool provides a rule-based segmentation and linguistic knowledge. Another useful tool is SPPAS (Bigi & Hirst 2012) to automatically produce annotations which include utterance, word, syllable and phoneme segmentation with support for

©Universitat de Barcelona

French, English, Italian and Chinese; output can be used later by other tools for prosodic analysis. In addition, the AuToBI Rosenberg (2009, 2010) tool for automatic generation of hypothesized ToBI labels, which has been applied to Portuguese (Moniz et al. 2014).

However, for many minority languages with a deficit of linguistic resources, such as Galician, a tool capable of automating these tasks has not yet been developed, sometimes due to the lack of acoustic models or to the lack of other resources such as grapheme-phoneme converters. Currently, specifically for Galician, there was no tools at user level that would allow to use the acoustic models obtained by the Multimedia Technology Group of the Universidade de Vigo for this language, combined with a grapheme-phoneme converter implemented by the same group mentioned before in collaboration with linguists on the project *Síntese de voz* of the Ramón Piñeiro centre and with one recogniser for performing an alignment. In this case, the recogniser used is provided by the HTK package. For this reason, the system board combines different resources in a single tool usable by any user without experience in speech processing; the diverse resources were mainly tailored to the needs of the AMPER project, because this is a large big database in different romance languages in which the data concerned follows a specific structure. The main novelty provided in this work is the development of an automatic tool which comprises many resources. Moreover, this programme is the first of its kind which includes segmentation and labelling for Galician, providing its users with an improved tool in terms of time spent on these tasks.

From this point of view, there are three important aspects which should be strengthened:

1. Automatic segmentation and vowel labelling, which will be the basis on which the prosodic analysis is sustained.

2. The integration of existing tools.

3. The tool is currently implemented to be employed primarily by researchers working within the framework of AMPER methodology.

In this way, in this paper we present a package that enables the alignment of audio and text devoid of timestamps, in addition to subsequent vowel labelling.

©Universitat de Barcelona

Furthermore, support for other languages with available voice recordings of appropriate quality or acoustic models has also been implemented: Galician, Spanish and Brazilian Portuguese. The addition of more languages is expected in the future, particularly the remaining Romance varieties, although it would be necessary to have acoustic models or quality speech databases and grapheme-phoneme converters in the language intended to be incorporated. This integration of a larger number of linguistic varieties would greatly increase the AMPER database and, therefore, allow a much more complete dialectal map of Romance intonation.

On the other hand, a second objective achieved with this work has been to seek the integration of all the tools most employed by the groups participating in the AMPER project for the extraction of F0 data, duration and energy, generation of graphics, etc., that provide a prosodic analysis of the recordings. In the future, the integration of R scripts which provide dialectrometric analysis (Martínez Calvo & Fernández Rei 2015) is also intended.

With this tool, named SEA AP, the automation of certain tasks that were previously performed manually, especially in data acquisition (measuring, labelling, transferring to an Excel sheet, etc.), is enabled. Thus we are witnessing an extraction and storage of data performed automatically and which will be later used for study and analysis. The tool is aimed at researchers who seek to focus their efforts on the analysis and interpretation of data, rather than on segmentation, labelling and obtainment of prosodic data.

The tool is freely available for use by the entire research community. To access the repository, the Multimedia Technology Group of the Universidade de Vigo, the Instituto da Lingua Galega of the Universidade de Santiago de Compostela, or any of the authors of this paper should be contacted.

This paper describes the details of the development and use of the tool. First, it provides a brief description of the tool explaining its characteristics and some aspects with regard to the internal operation. Then, there is an explanation of the module in charge of the segmentation and labelling of the input material. In the following section, the scripts used for the prosodic analysis integrated there are described.

©Universitat de Barcelona

Finally, the conclusions obtained from this work and the resulting future lines are set out.

2. Tool features

SEA_AP (***S**egmentador e **E**tiquetador **A**utomático para **A**nálise **P**rosódica, Automatic Segmentation and Labelling for Prosodic Analysis*) is a desktop application, compatible with the Windows operating system, allowing the alignment of audio files with their respective transcription, as well as their prosodic analysis. The alignment results produces segmentation with temporal labelling with regard to words, syllables and/or phonemes. In the main window of the application, the type of units for the segmentation, as well as the language to analyse the data, can be selected.

After this first phase of alignment, the application generates the data needed for a prosodic analysis using the Praat software; it will therefore be run along with integrated scripts performing interactions with the user so that he/she can modify certain parameters used in the analysis, if desired, as well as the visualization of the results obtained from the analysis executed.

The data that will be provided for the tool is the directory where the audio files are, the directory with the texts to be aligned, the directory where we would like the TextGrid for Praat to be generated and, finally, the directory for storing the results from the prosodic analysis. All this information is required for the proper operation of the tool.

Once all this input info has been entered, the segmentation and labelling tasks will start when the alignment button is clicked. In order to carry them out, the first step performed is an automatic phonetic transcription of a given text in .txt format, transforming the graphemes in phonemes. Subsequently, a forced recognition assigning the aforementioned units to audio segment by means of the recognition engine HTK (Hidden Markov Model Toolkit) (Young et al. 1995) will be performed. Finally, the scripts for the prosodic study will be executed based on the previous segmentation and labelling.

©Universitat de Barcelona

3. Elements of the alignment block

This section will deal with a brief description of the different elements that make the alignment of the sound waveform and the orthographic transcription provided possible.

Firstly, for the automatic phonetic transcription for Galician and Spanish languages, the Cotovia package has been used (González 2004, Rodríguez Banga et al. 2012), which was developed by the centre Ramón Piñeiro in collaboration with the Universidade de Vigo and the Universidade de Santiago de Compostela, groups to which the authors belong. The system used for Portuguese has been obtained from the FalaBrasil project (Nelson et al. 2010), developed by the Universidade Federal do Pará.

Secondly, the system for labelling phonemes is implemented by using the HTK tool, consisting of a series of modules that perform the necessary tasks for the implementation of recognisers based on HMM (Torre & Hernández 2002).

3.1 Automatic phonetic transcription

As mentioned above, the first step is to obtain an automatic phonetic transcription (González et al. 2008) of the text to be aligned with the sound waveform. For this task, a number of steps are performed:

a) Linguistic Preprocessing: the texts usually contain words such as acronyms or numbers that needs to be normalised so that the conversion is correctly completed. For this, the following steps are carried out:

- An expansion of the abbreviations and acronyms contained in the sequence.

- The replacement of numbers, both Arabic and Roman, using certain reading rules, as well as the transcription of time expressions.

- The transcription of foreign words by adjusting the sounds used in the language studied, since as a general rule grapheme-sound correspondence is particular to each language.

©Universitat de Barcelona

b) A morphological analysis: this involves the classification of words according to their grammatical category, gender and number as described in Seijo et al. (2004). The steps to be taken to perform this classification are the following:

- The labelling of words belonging to grammatical groups with low variability, such as prepositions, adverbs, conjunctions or determinants.

- An analysis of the roots and morphemes of words with greater variability, such as names, adjectives or verbs, by assigning them all possible categories from amongst which the definitive one will be chosen afterwards. This disambiguation is carried out by means of a contextual analysis which takes into account the categories that were already assigned to the preceding and following words. In addition, a statistical model which takes into consideration the most probable category is also applied.

- The study of word groups such as adverbial or conjunctive phrases and verbal periphrasis.

After categories are assigned, words are divided into syllables by using an algorithm, implemented by the Cotovia package (González 2004, Rodríguez Banga et al. 2012), which detects the syllable nucleus (which is always a vowel in the case of the languages supported by the application) and which takes into account the surrounding consonants in order to determine the syllable boundaries. It is also at this point when stress is assigned. The syllabication algorithm takes into consideration diphthongs and triphthongs and allows the detection of less frequent phoneme sequences that appear in words from Latin, Greek or other foreign languages.

c) A syntactic analysis: this step is focused on the analysis of the sentences taking into account the existing prepositions and the way in which the words and syntagms relate to each other. The position of punctuation signs and words such as conjunctions, conjunctive phrases or relative pronouns is also taken into account.

It also takes into account the position of punctuation and words such as conjunctions, conjunctive phrases or relative pronouns.

In addition, a re-accentuation of words is accomplished by removing the stress of function words and by inserting pauses. The latter are inserted through to punctuation signs and in sequences exceeding a certain length.

©Universitat de Barcelona

d) A phonetic transcription: In this phase the grapheme string becomes a sequence of phonetic characters representing the allophones.

3.2 Automatic Labelling of Phonemes

Once the allophonic representation of the text is available, the next step will be to obtain their occurrence temporal instants, for this purpose an automatic labelling is performed.

A block diagram of the system used is shown below:

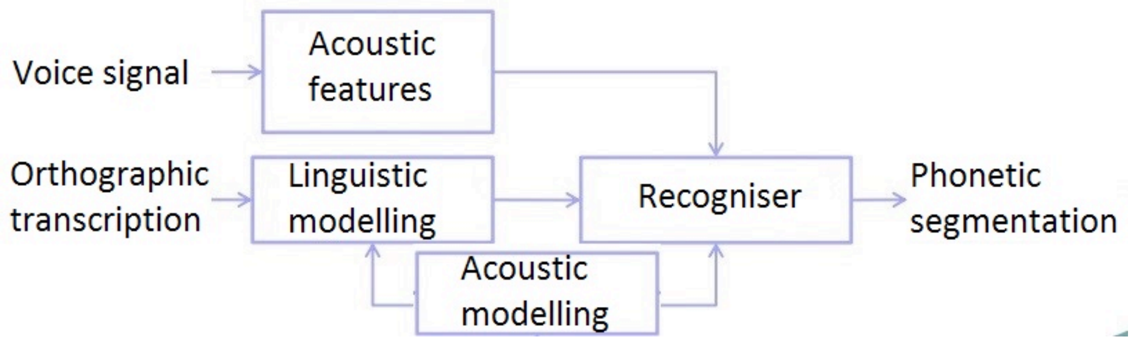


Figure 1. Diagram of the alignment module

As shown, the system mainly consists of four modules which are described below:

a) Acoustic features module: the system takes an audio signal as input, and this module is responsible for extracting the features which intend to model the behaviour that the human ear would display in the presence of a similar stimulus. In this case, the features used are the *Mel-Frequency Cepstral Coefficients, MFCCs.* These coefficients are suitable given that they adequately simulate the human auditory perception and, furthermore, they are robust against background noise.

b) Acoustic modelling module: this module is the responsible for relating certain sound waveforms with the corresponding allophone, with each allophone therefore represented as a sound. However, a phoneme has many variations given that one person does not always pronounce a word in the same way, different people have

©Universitat de Barcelona

different pronunciations, etc. For all this, this task is a major problem and it has been the object of numerous studies.

It has been proved that good results can be achieved using those known as the *Hidden Markov Models, HMMs* (Rabiner 1989). The purpose of these models is the numeric representation of each of the phonemes in question, because as mentioned above, they could vary. Thus, these models are for gathering information from different sources and obtaining the representation that suits the greatest number of individuals. These models are obtained through inductive learning, which means they gain knowledge from previously-known examples. The generation of these models is performed by developments carried out using large amounts of labelled data. The greater the quantity and quality of acoustic data used for the training phase, the greater accuracy such models will have and, therefore, the better the final result obtained in the recognition step will be.

The acoustic models used for the forced recognition in Spanish and Galician have been developed by atlantTIC research group with the following material:

- Spanish: developed with part of the TCSTAR database (Docio-Fernández et al. 2006) which contains recordings from the European and Spanish parliaments, labelled with their corresponding transcriptions.

- Galician: developed with the Transcrigal database (Garcia-Mateo et al. 2004), containing 31 hours recordings of news broadcast on *Televisión de Galicia* (TVG).

We have decided to use this material and not AMPER for several reasons. On the one hand, this material covers different scenarios and contains several speakers, who usually speak planned speech, and there are also other speakers speaking spontaneous speech. Different accents and dialectal variations appear. In addition, there are long sentences, resulting in a greater number of triphones. On the other hand, the AMPER database includes a reduced number of stages, and although it comprises many speakers, all of these employ planned speech. This corpus only contains small sentences of around 11-14 syllables and the the same structure. In addition, this database has an insufficient number of hours to obtain robust acoustic models, and therefore adaptation to this type of material is avoided.

©Universitat de Barcelona

In the case of Brazilian Portuguese, the models available in the repository of the FalaBrasil project (Nelson et al. 2010), that were already trained by their authors, have been used

c) Linguistic modelling module: this is responsible for creating the network that defines the structure of the sentence. In large vocabulary applications, statistical models are typically used, from which the most likely word is predicted. Nevertheless, in this case, as the sequence of words is previously known, a tree is built that forces the assignment of words to each of the temporal segments.

d) Recognition Module: this is tasked with combining the information from the previous module and assigning a time interval to each of the segments taken into consideration, as well as a certain probability. This process is performed by means of what is known as forced recognition, by applying the Viterbi algorithm. It is called 'forced' because instead of enabling the system to freely choose the word, it is forced to select the word corresponding to the sequence provided by the linguistic module.

4. Features of the prosodic module

The study of the phonic features is carried out by integrating in the tool different Praat scripts, mainly implemented by the team of the Phonetics Laboratory of the Universitat de Barcelona (Martínez Celdrán & Fernández Planas, 2003-2015). The Praat scripts from the above mentioned team that have been integrated are the following:

- create_pictures.praat (Elvira-García & Roseano 2014): this script allows the creation and storage of images associated with the audio files included in a folder. The images can contain the waveform, the spectrogram, a curve of evolution of the fundamental frequency (F0), and the content of the different lines of the TextGrid associated with each of the sound files. Therefore, this script is very useful when representing and storing the audio waveform considered along with the automatically obtained TextGrid which has been subjected to manual review.

Below, a figure is shown with the result of one of the studies carried out, in this case for Galician.

©Universitat de Barcelona

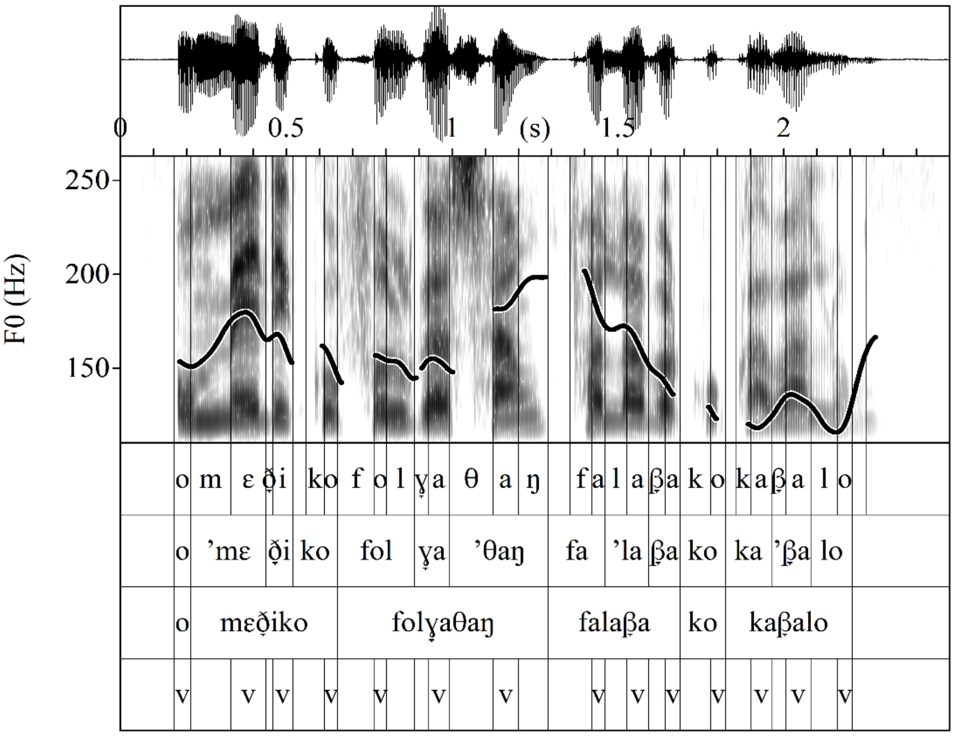


Figure 2. Results of the script create_pictures.praat

- Prosodic_data_extraction.praat (Elvira-García 2014a): this generates a text file that can contain information regarding duration, intensity and fundamental frequency in Hertz or semitones for three points of each non-empty interval of the TextGrid. Furthermore, it extracts the labels from those intervals or points and the position of the data with regard to the stressed vowel.

- Blank_TextGrid_creation.praat (Elvira-García 2014b): when given a particular folder with audio files, this script is used to generate empty TextGrid files for each of them.

- Remove_tiers.praat (Elvira-García 2014c): this is used to automatically delete tiers in all the TextGrid of a particular folder.

- Created_TextGrid_modification.praat (Elvira-García 2014d): this is used to make corrections in an existing TextGrid allowing the user to visualize it along with its corresponding audio and edit it.

©Universitat de Barcelona

In addition to these, the following script, created at the Universidade de Vigo, has been integrated with the aim of reducing the time of manual review by the user, thus improving the tool´s performance.

- Regenera_textgrid.praat: it allows to generate, from a TextGrid containing only the phonemes tier, the divisions for a syllable tier and other of words. Thus, it is intended that the user should only make corrections in the phonetic tier, and that those changes could be automatically propagated to the other syllable and word tier. In addition, it is responsible for producing an extra tier containing information relating to the segmentation of the vowels.

The AMPER_PRAAT_Textgrid2Txt_V3.praat (Rilliard 2013) script, widely used by all the members of the AMPER Project, has also been added. This script allows the logic implemented in Praat to extract, from vowel segmentation, prosodic parameters such as fundamental frequency (F0), duration and intensity, to be used. Once this data is obtained, it can be stored in a text file in AMPER format, over the programme models in Matlab written by Antonio Romano. This script is based on ExtF0forVowels (Barbosa 2006) and has been improved with a functionality which enables users to correct the F0 contour manually, in order to correct any algorithm errors and obtain more accurate information in the output; also, the output to the script is stored in a TXT file in AMPER format.

5. Example of use

SEA_AP has been designed to maximise the user experience by means of the automation of every possible task. In this section, we will present a brief user guide in order to facilitate the understanding of its function.

Suppose that we have a corpus of audio with its respective transcriptions whose labelling would be carried out through SEA_AP. First, the preparation of the data, audio and transcription files which may be located in the same directory, the same folder for audio and transcription, or in separate directories, one for audios and another one for transcriptions, is required. Furthermore, the names assigned to the

©Universitat de Barcelona

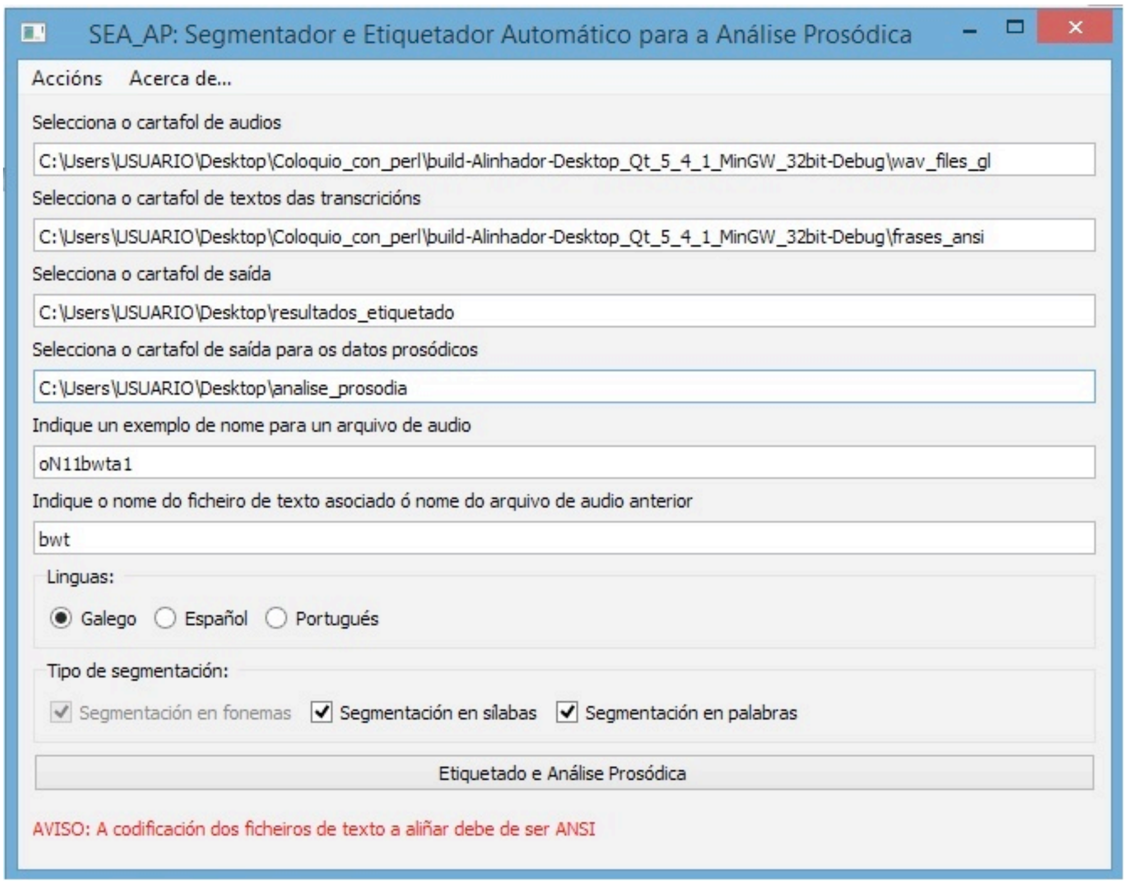
audios and to the transcriptions have to follow a specific pattern, so that the system can associate each audio with its transcript following the calculated pattern. For instance, if we have an audio identified by oK11**bwk**a1.wav and its transcript by **bwk**.txt, the system will determine that the name of the transcription file starts at position 5 of the name of the audio and that it considers three letters from that position. In order for the tool to calculate this pattern, the user must enter an example of an audio name and another of a transcription name. In the event that both the audio and the transcript have the same name, it is not necessary to fill in those fields.

Once we have the material prepared, it is sufficient to fill the form shown in Figure 3, where it is required to specify the directories where the audio and transcription files are. The directories where we would like to save the data generated by the tool must then be specified: on the one hand, the directory where the labelling results will be stored (TextGrid files which can be later used by Praat software); and, on the other hand, the directory where the files concerning the results of the prosodic analysis will be stored, whether images or text files, with data relating to fundamental frequency, duration, etc. Following this, an example of a name assigned to an audio and an example of the name assigned to a file containing the related transcription is required, so that the system calculates the relation of these names to each other, as it obtains the pattern followed by the names, so all the names must therefore have the same structure. For example, if we have audios with names 01**aa**23.wav and 02**aa**99.wav and a file transcription named **aa**.txt, both audio files will be related to this transcription file.

Afterwards, the language of the audios to be aligned and segmented must be indicated. Additionally, the type of segmentation desired needs to be specified: in syllables and/or words. The option of phonetic segmentation must be always performed, therefore the users are not allowed to deactivate this checkbox.

Finally, the alignment button must be pressed in order to start the process of automatic segmentation and labelling. This process is transparent to the user, and it allows the TextGrid files linked to the alignment of the audios with the text to be obtained, in the directory we have indicated on the form.

©Universitat de Barcelona


SEA_AP: Segmentador e Etiquetador Automático para a Análise Prosódica

Selecciona o cartafol de audios
C:\Users\USUARIO\Desktop\Coloquio_con_perl\build-Alinhador-Desktop_Qt_5_4_1_MinGW_32bit-Debug\wav_files_gl

Selecciona o cartafol de textos das transcricións
C:\Users\USUARIO\Desktop\Coloquio_con_perl\build-Alinhador-Desktop_Qt_5_4_1_MinGW_32bit-Debug\frases_ansi

Selecciona o cartafol de saída
C:\Users\USUARIO\Desktop\resultados_etiquetado

Selecciona o cartafol de saída para os datos prosódicos
C:\Users\USUARIO\Desktop\analise_prosodia

Indique un exemplo de nome para un arquivo de audio
oN11bwta1

Indique o nome do ficheiro de texto asociado ó nome do arquivo de audio anterior
bwt

Linguas:
● Galego ○ Español ○ Portugués

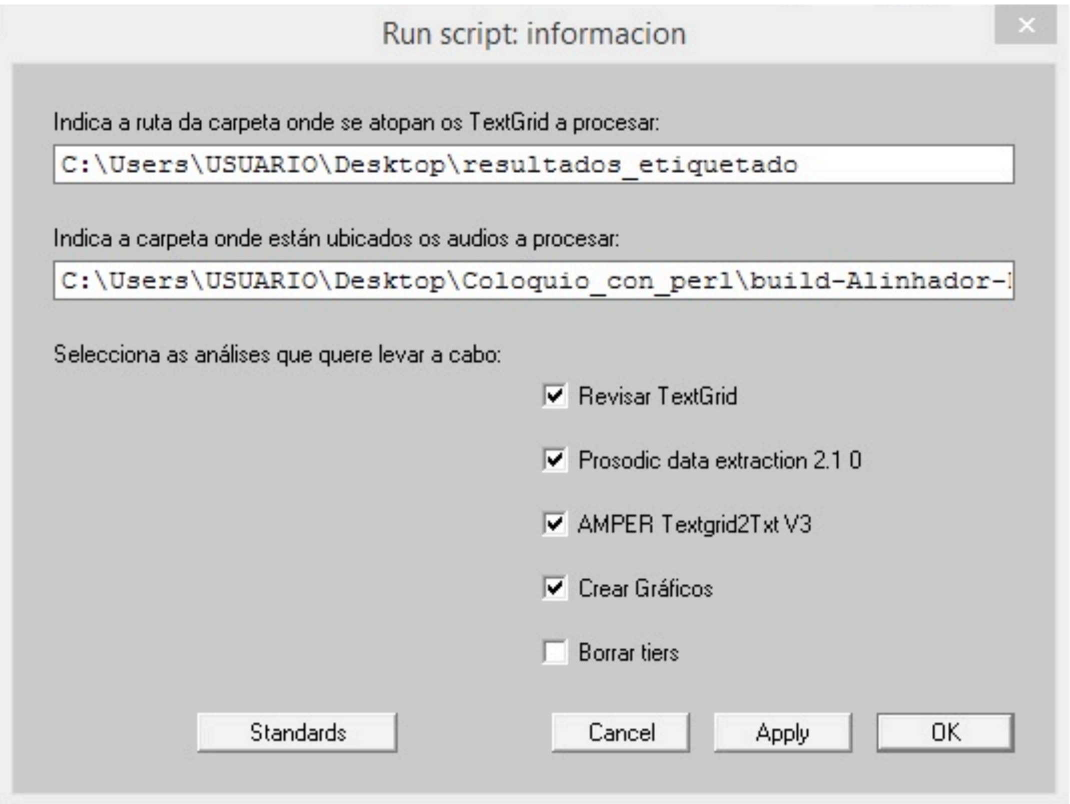
Tipo de segmentación:
☑ Segmentación en fonemas ☑ Segmentación en sílabas ☑ Segmentación en palabras

AVISO: A codificación dos ficheiros de texto a aliñar debe de ser ANSI

Figure 3. Data entry form. Interface language is Galician

Subsequently, the system automatically opens the Praat software together with the script *principal.praat* that was designed in order to automatically generate the form that will be displayed to the user, where it can be selected everything related to prosodic analysis. The system fills fields with data already entered by the user, but they can be modified if desired. Figure 4 shows the form produced by the script above mentioned.

©Universitat de Barcelona



Indica a ruta da carpeta onde se atopan os TextGrid a procesar:
C:\Users\USUARIO\Desktop\resultados_etiquetado

Indica a carpeta onde están ubicados os audios a procesar:
C:\Users\USUARIO\Desktop\Coloquio_con_perl\build-Alinhador-

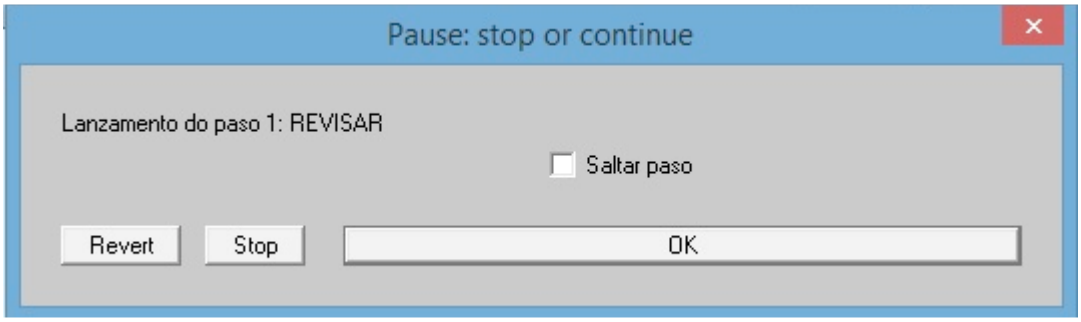
Selecciona as análises que quere levar a cabo:

- ☑ Revisar TextGrid
- ☑ Prosodic data extraction 2.1 0
- ☑ AMPER Textgrid2Txt V3
- ☑ Crear Gráficos
- □ Borrar tiers

Figure 4. Data entry form for prosodic analysis

In this form, as mentioned above, the system automatically fills the fields of the directories where the audio files and the associated TextGrid are located, with the data entered by the user in the first form. Moreover, the user must select the types of analysis desired, either checking or unchecking the boxes provided for that purpose.

Once the user clicks OK on the form, the scripts selected will begin to run. In any case, the user is informed about the step to be executed and provided with the possibility to skip it, if desired, running then the following step of the list. (Figure 5).



Lanzamento do paso 1: REVISAR

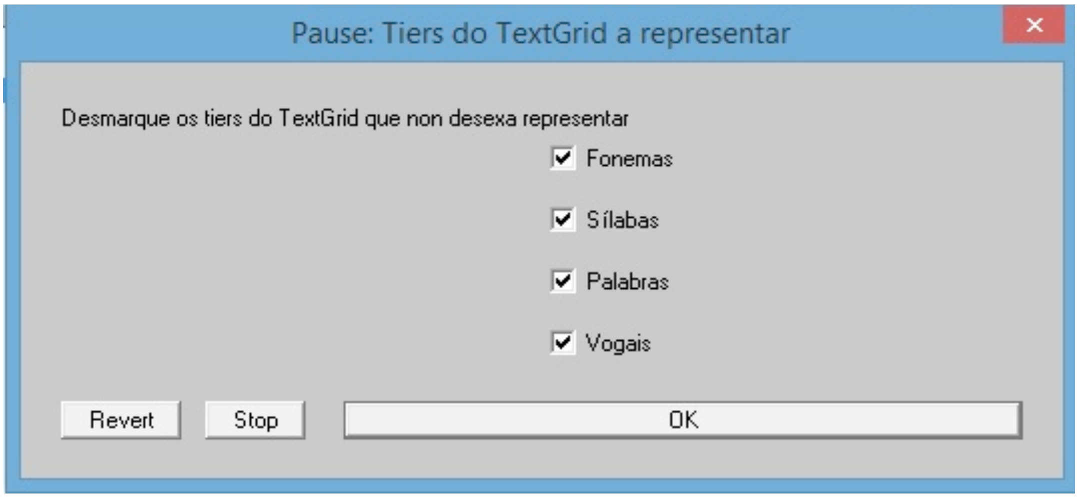
□ Saltar paso

Figure 5. Information window displayed to the user before each step

©Universitat de Barcelona

For the review task, the TextGrid generated in the first step is shown, which only contain the phonetic segmentation. Thus, at this stage of manual review, the user only has to correct that layer, from which, once corrected, the system will generate the rest of the lines. In the event that the user does not want to launch this step, the system will generate the remaining lines with the available segmentation in syllables and words.

In the task of "Graph Generation", the form is shown in Figure 6, where the user can select the tiers that will appear on the graph generated from the TextGrid, avoiding the display of those not considered interesting for their purposes.



Pause: Tiers do TextGrid a representar

Desmarque os tiers do TextGrid que non desexa representar

☑ Fonemas

☑ Sílabas

☑ Palabras

☑ Vogais

Figure 6. Window for the selection of tiers to be shown in the graph within the TextGrid

The remaining steps perform the tasks described in the scripts of the previous section.

6. Conclusions

The work presented in this project has required the collaboration of numerous research sources from different areas carried out at different universities, seeking to attain the ultimate objective of integrating of all the improvements provided by each of them in a single robust and easy to use tool.

©Universitat de Barcelona

Therefore, based on the needs expressed by the specialists on prosodic analysis, a search has been performed of the most useful tools which are the result of several research activities, and that, grouped together, constitute a very beneficial contribution in order to automate certain works that were otherwise very laborious.

Finally, as a result of the cooperation between the engineering and linguistic areas, and thanks also to the improvements made, we have been able to automate the time-consuming manual segmentation and labelling by means of an easy-to-use desktop application, tailored to the needs of the target users, which allows a considerable saving on the time spent. Furthermore, it brings together segmentation, labelling and prosodic analysis in a single tool.

7. Future work

We are dealing with a complete and ready to use tool. However, as a continuation of this project, we may propose to undertake a case study to evaluate the performance of the alignment by means of a comparison between a manually labelled database and a labelling performed with SEA_AP.

Given that the performance of the alignment largely relies on the quality of the forced recognition and, therefore, on the acoustic models, it is also possible to improve those models by adapting them to the material aligned using this tool. Thus, the quality of the next alignment and the existing repositories for a particular language in the minority languages would be improved.

In addition, this tool could be useful in less controlled scenarios than those originally conceived, since it could be applied in spoken language corpus of different levels and registers: rural language, urban language, formal language, etc. (Escourido et al. 2008). An example of this type of corpus would be CORILGA: *Corpus Oral Informatizado de la Lengua Gallega* (García-Mateo et al. 2014). This way, a prosodic labelling of the material collected in this corpus could be provided, which would be a significant advance as it would allow us to address the study of intonation in

©Universitat de Barcelona

spontaneous interactions, one of the major challenges in current research in the field of prosodic studies.

Another potential line of research is the application of a methodology of dialectometric analysis to the prosodic data generated by our tool. To the extent that SEA_AP allows the labelling and analysis of large quantities of sequences, the integration of scripts for the dialectometric analysis of the results (Martínez Calvo & Fernández Rei 2015) could be proposed.

Finally, it is worth noting the possibility of expanding this to other languages with sufficient data available in order to train the acoustic model and a suitable grapheme-phoneme converter, and to also create a multilingual user interface.

8. Acknowledgements

This work would have not been possible without the help of the Spanish Government (Project ‘SpeechTech4All’ TEC2012-38939-C03-01), the European Regional Development Fund (ERDF), the Government of the Autonomous Community of Galicia (GRC2014/024, “Consolidación de Unidades de Investigación: Proyecto AtlantTIC” CN2012/160) and the “Red de Investigación TecAnDAli” from the Council of Culture, Education and University Planning, Xunta de Galicia.

We would also like to thank the anonymous reviewers for their suggestions which contributed to improving the article.

References

BARBOSA, P.A. (2006) *Incursões em torno do ritmo da fala*, Campinas: Pontes.

BIGI, B. & D. HIRST (2012) “Speech Phonetization Alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody”, *Proceedings of Speech Prosody*, Shanghai, 1-4. <https://hal.archives-ouvertes.fr/hal-00983699>

©Universitat de Barcelona

BOERSMA, P. & D. WEENIK (2013) *Praat: doing phonetics by computer* [Computer program], version 5.4.05 [http://www.praat.org/].

CONTINI, M., J.P. LAI & A. ROMANO (2002) "La géolinguistique à Grenoble: de l'AliR à l'AMPER", in M.R. Simoni-Aurembou (ed.), *Nouveaux regards sur la variation diatopique, Revue belge de Philologie e d'Histoire*, 80, 931-941.

DOCIO-FERNÁNDEZ, L., A. CARDENAL-LÓPEZ, C. GARCÍA-MATEO (2006) "TC-STAR 2006 Automatic Speech Recognition Evaluation: The UVIGO System", in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, 145-150. <https://www.researchgate.net/publication/255661612_TC-STAR_2006_Automatic_Speech_Recognition_Evaluation_The_UVIGO_System>

ELVIRA-GARCÍA, W. (2014a) *Prosodic-data-extraction v2.1* [Praat script, distributed under GNU General Public License] <http://stel.ub.edu/labfon/en/praat-scripts>

ELVIRA-GARCÍA, W. (2014b) *Blank_TextGrid_creation* [Praat script, distributed under GNU General Public License] <http://stel.ub.edu/labfon/en/praat-scripts>

ELVIRA-GARCÍA, W. (2014c) *Remove_tiers* [Praat script, distributed under GNU General Public License] <http://stel.ub.edu/labfon/en/praat-scripts>

ELVIRA-GARCÍA, W. (2014d) *Created_TextGrid_modification* [Praat script, distributed under GNU General Public License] <http://stel.ub.edu/labfon/en/praat-scripts>

ELVIRA-GARCÍA, W. & P. ROSEANO (2014) *Create pictures with tiers v.4.1.* [Praat script, distributed under GNU General Public License] <http://stel.ub.edu/labfon/en/praat-scripts>

ESCOURIDO, A., E. FERNÁNDEZ REI, M. GONZÁLEZ & X. L. REGUEIRA (2008) "A dimensión prosódica da oralidade. Achega dende AMPER", in E. Fernández Rei & X.L. Regueira, *Perspectivas sobre a oralidade*, Santiago de Compostela: Instituto da Lingua Galega / Consello da Cultura Galega, 75-93.

FERNÁNDEZ PLANAS, A.M., P. ROSEANO, E. MARTÍNEZ-CELDRÁN & L. ROMERA (2011) "Aproximación al análisis dialectométrico de la entonación en algunos puntos del dominio lingüístico catalán", *Estudios de Fonética Experimental*, XX, 141-178.

FERNÁNDEZ PLANAS, A.M., J. DORTA, P. ROSEANO, Ch. DÍAZ, W. ELVIRA-GARCÍA & E. MARTÍNEZ-CELDRÁN (2015) "Distancia y proximidad prosódica entre algunas variedades del español: un estudio dialectométrico a partir de datos acústicos", *Revista de Lingüística Teórica y Aplicada*,53 (2), 13-45.

FROTA, S. & P. PRIETO (eds.) (2015) *Intonational Variation in Romance*, Oxford: Oxford University Press.

©Universitat de Barcelona

GARCIA-MATEO, C., J. DIEGUEZ-TIRADO, A. CARDENAL-LOPEZ, & L. DOCIO-FERNANDEZ (2004) "Transcrigal: A bilingual system for automatic indexing of broadcast news", in *Proceedings Int. Conf. on Language Resources and Evaluation*, volume 6, Lisbon: ELRA, European Language Resources Association, 2061-2064. <http://www.lrec-conf.org/proceedings/lrec2004/pdf/382.pdf>

GARCÍA-MATEO, C., A. CARDENAL, X.L. REGUEIRA FERNÁNDEZ, E. FERNÁNDEZ REI, M. MARTÍNEZ, R. SEARA, R. VARELA & N. BASANTA LLANES (2014) "CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis", in N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik: ELRA <http://www.lrec-conf.org/proceedings/lrec2014/index.html>

GARRIDO, J. M. (2013) "SegProso: A Praat-Based Tool for the Automatic Detection and Annotation of Prosodic Boundaries", *Proceedings of TRASP*, 74-77. <http://www.lpl-aix.fr/~trasp/Proceedings/19864-trasp2013.pdf>

GOLDMAN, J. P. (2011) "EasyAlign: an automatic phonetic alignment tool under Praat", *Proceedings of InterSpeech*, Firenze, Italy, 3233-3236. <http://latlcui.unige.ch/phonetique/easyalign/easyalign_unpublished.pdf>

GONZÁLEZ, M. (2004) "A síntese de voz en lingua galega: O proxecto Cotovía", *Revista galega do ensino*, 44, 199-215.

GONZÁLEZ GONZÁLEZ, M., E. RODRÍGUEZ BANGA, F. CAMPILLO DÍAZ, F. MÉNDEZ PAZÓ, L. RODRÍGUEZ LIÑARES & G. IGLESIAS IGLESIAS (2008) "Specific features of the Galician language and implications for speech technology development", *Speech Communication*, 50, 874-887.

MAIRANO, P. (ed.) (2011) *Intonations Romanes*, *Géolinguistique*, hors-série 4.

MARTÍNEZ CALVO, A. & E. FERNÁNDEZ REI (2015) "Unha ferramenta informática para a análise dialectométrica da prosodia", *Estudios de Fonética Experimental*, XXIV, 289-303 <http://stel.ub.edu/labfon/sites/default/files/9_MARTINEZ.pdf>

MARTÍNEZ CELDRÁN, E. & A. M. FERNÁNDEZ PLANAS, (coords.) (2003-2015) *Atlas Multimèdia de la Prosòdia de l'Espai Romànic*. <http://stel.ub.edu/labfon/amper/cast/index_ampercat.html>

MONIZ, H., A.I. MATA, J. HIRSCHBERG, F. BATISTA, A. ROSENBERG & I. TRANCOSO (2014) "Extending AuToBI to prominence detection in European Portuguese", In N. Campbell, D. Gibbon & D. Hirst (eds.), *Proceedings of Speech Prosody*, Dublin, Trinity College, 280-284 <http://www.speechprosody2014.org/>

©Universitat de Barcelona

MOUTINHO, L.C., R.L. COIMBRA, A. RILLIARD & A. ROMANO (2011) “Mesure de la variation prosodique diatopique en portugais européen”, *Estudios de Fonética Experimental*, 20, 33-55.

NELSON NETO, P. SILVA, A. KLAUTAU & I. TRANCOSO (2010) “Free tools and resources for Brazilian Portuguese speech recognition”, *Journal of the Brazilian Computer Society* <http://link.springer.com/article/10.1007%2Fs13173-010-0023-1>

RABINER, L.R. (1989) “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proceedings of the IEEE*, Vol. 77, No. 2, 257-286. <http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf>

RILLIARD, A. (2013) “Metodoloxía cuantitativa para a medida das distancias prosódicas”, *Xornadas de Dialectoloxía Perceptiva*, Santiago de Compostela <http://ilg.usc.es/tecandali/Descargas/AlbertRilliard.pdf>

RODRÍGUEZ BANGA, E, C. GARCÍA MATEO, F. J. MÉNDEZ PAZÓ, M. GONZÁLEZ & C. MAGARIÑOS IGLESIAS (2012) “Cotovía: an open source TTS for Galician and Spanish”, VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, *IberSPEECH 2012*, Madrid <http://iberspeech2012.ii.uam.es/IberSPEECH2012_OnlineProceedings.pdf>

ROSENBERG, A. (2009) “Automatic detection and classification of prosodic events”, Columbia University, Ph.D. Thesis. <http://www1.cs.columbia.edu/~amaxwell/amaxwell-thesis-final.pdf>

ROSENBERG, A. (2010) “AuToBI-A tool for automatic ToBI annotation”, in INTERSPEECH, 146-149. <http://eniac.cs.qc.cuny.edu/andrew/papers/autobi-is10.pdf>

SEIJO PEREIRO, L., A. MARTÍNEZ ÍNSUA, F. MÉNDEZ PAZÓ, F. CAMPILLO DÍAZ & E. RODRÍGUEZ BANGA (2004) “A Galician Textual Corpus for Morphosyntactic Tagging with Application to Text-to-Speech Synthesis”, in *Proceeding of LREC* 2004, Lisboa, vol. 5, 1759-1762. <http://www.lrec-conf.org/proceedings/lrec2004/pdf/111.pdf>

TORRE TOLEDANO, D. & L. HERNÁNDEZ GÓMEZ (2002) “Hmms for automatic phonetic segmentation”, in *Proc. of LREC* 2002, Las Palmas de Gran Canaria [http://www.lrec-conf.org/proceedings/lrec2002/].

YOUNG, S., G. EVERMANN, M. GALES, T. HAIN, D. KERSHAW, G. MOORE, J. ODELL, D. OLLASON, D. POVEY, V. VALTCHEV & P. WOODLAND (1995) *HTK Book*, University of Cambridge <http://htk.eng.cam.ac.uk/docs/docs.shtml>

©Universitat de Barcelona