EasyAlign: an (semi-) automatic segmentation tool under Praat for Brazilian Portuguese

Jean-Philippe GOLDMAN (Université de Genève)
Maíra AVELAR MIRANDA (PUC-MG)
Cirineu CECOTE STEIN (UFPB)
Antoine AUCHLIN (Université de Genève)

The purpose of phonetic alignment (or phonetic segmentation) is to determine the time position of phone boundaries in a speech corpus on the basis of the audio recording and its orthographic transcription. Aligned corpora are widely used in various speech applications like automatic speech recognition, speech synthesis, as well as prosodic and phonetic research. Although manual segmentation constitutes the more accurate method, it requires a large amount of time and concentration for the human labeller. Thus, various automatic methods are now used as they are not only much quicker, but their results are also reproducible and consistent throughout a large corpus. Unfortunately, whatever the automatic tool chosen, computational skills are often needed and many steps are required to prepare the data before the automatic tool can do its job.

EasyAlign has been developed in order to provide an ergonomic automatic segmentation tool, easy to use for computer science non-specialists. Although it is HTK-based [1], its topmost layer, i.e. the user interface, lies within Praat software [2], which hides the in-line commands that require non-trivial computational skills. Moreover, it is distributed as a self-installable plug-in, and its tools are directly accessible from the Praat menus. It is available for French, English, Castilian Spanish and recently for Brazilian Portuguese.

EasyAlign consists in a group of tools which successively perform three processes: utterance segmentation, grapheme-to-phoneme conversion and phonetic segmentation. Given that recognition tools are not designed to process unlimited-length utterances, the automatic phonetic alignment process needs a first segmentation into utterances. From the orthographic transcription (one utterance per line within a text file), the utterance segmentation process – which is language-independent– generates a TextGrid with one tier, in which each interval contains one utterance. Then, the grapheme-phoneme conversion step creates a tier with the phonetic transcription of the utterances. For this language-specific process, we used for Brazilian Portuguese the phonetizer developed by the Fala Brasil team [3], which provides a detailed phonetic transcription in SAMPA. Some adjustments have been made in the original dictionary [4] used to convert the words, especially concerning the exceptions related to the open vowels [E] and [O], which were not converted properly in some contexts. In order to solve this problem, some word corrections have been made manually. Furthermore, the number of words with the open vowels [E] and [O] were manually expanded: at first the words were selected from the Aurélio dictionary (electronic version) [5], then, the words were converted and, finally, a manual correction of the vowels [E] and [O] was made. Consequently, we used the Fala Brasil phonetizer with corrected and expanded dictionary.

Finally, in the phonetic segmentation step, the Viterbi-based HVite tool (within HTK) is called to align each utterance to its verified phonetic sequence. For Brazilian Portuguese, the acoustic models were trained on the basis of about 20 minutes of unaligned multi-speaker speech for which a verified phonetic transcription was provided. The phonetic segmentation process generates one tier with phones and another one with words. Additionally, a syllable tier is generated on the basis of sonority-based rules for syllable segmentation. In summary, the whole process results, with some minor manual verifications and adjustments to ensure better quality, in a multi-level annotation within a TextGrid composed of phonetic, syllabic, lexical and

utterance tiers. The evaluation of EasyAlign for Brazilian Portuguese is currently in progress, whereas the performances for French were fairly comparable to human segmentation.

In order to demonstrate the applicability of Easy Align, an analysis of ten minutes of speech will be performed. The samples will be extracted from a Brazilian political debate, broadcasted by Record in the second tour of the 2010 presidential elections: Five minutes of speech belonging to the candidate Serra and five minutes belonging to the candidate Dilma. A manual syllabic segmentation of the speech has already been made. So, we intend to compare the manual segmentation with the segmentation made automatically by Easy Align, as well as the two different prosodic profiles generated from these two types of segmentation.

In conclusion, EasyAlign is a friendly tool which enables to easily align speech from an orthographic transcription. To our knowledge, such a tool was not, until now, freely available for Brazilian Portuguese. This kind of tool can be very useful for the Brazilian scientific community, especially for the researchers that have to deal with a large amount of acoustic data.

References

[1] Young, S. et al. (2010). The HTK book. http://htk.eng.cam.ac.uk/. Last access: March 2010.
[2] Boersma, P. & Weenink, D. (2010). Praat: doing phonetics by computer.
http://www.praat.org. Last access March 2010.
[3] Klautau, A. et al (2011) Conversor Grafema-fone v1.6
http://www.laps.ufpa.br/falabrasil/downloads.php Last access: August 2011
[4] Klautau, A. et al (2010) UFPAdic 3.0 (2010)
http://www.laps.ufpa.br/falabrasil/downloads.php Last access: August 2011
[5] Dicionário Aurélio Eletrônico- Século XXI (1999) Rio de Janeiro: Nova Fronteira e Lexicon Informática, CD-rom, versão 3.0.