An Automatic Phonetic Aligner for Brazilian Portuguese with a Praat Interface

Gleidson Souza and Nelson Neto(✉)

Institute of Exact and Natural Sciences, Federal University of Pará,
Augusto Correa. 1, Belém, PA 660750110, Brazil
gleidson.sousa@itec.ufpa.br, nelsonneto@ufpa.br

Abstract. The analysis of the phonetic entities of speech nearly always requires the alignment of an audio file with its phonetic transcription. However, it is an extremely labor-intensive task. An automatic alignment tool has modules that depend on the language and, while there are many public resources for some languages (e.g., English and French), the resources for Brazilian Portuguese (BP) are still limited. This work describes the development of an automatic phonetic alignment tool for BP, consisting of grapheme-to-phone converter, syllabification system and HTK-based acoustic models. This aligner is implemented and freely distributed as a plug-in of Praat. Performance tests are presented, comparing the current proposal with an existing tool.

AQ1

Keywords: Phonetic alignment · Brazilian Portuguese · Pronunciation dictionary · Syllabification · HTK · Praat

1 Introduction

Automatic speech recognition (ASR) and speech synthesis (TTS) are data-driven technologies that require a relatively large amount of labeled data. As consequence, many large speech corpora have been collected for speech technology development in the recent years. And they need to be phonetically segmented with a high level of precision, i.e. the phones must be time-aligned with the sound, on risk of impairing the quality of the synthesized voice, for example. Indeed, the analysis of the prosodic structure of speech requires to know the precise position of the phonetic temporal boundaries [1]. However, manual phonetic segmentation is time-consuming, more than 13 h for a one-minute recording [2], and expensive, since it requires trained language experts.

The most widely explored phonetic alignment techniques are based either on hidden Markov models (HMM) used in forced-alignment mode or on dynamic time alignment with synthesized speech (TTS+DTW) [3]. In [4], a comparison between these two approaches has showed that in general the TTS+DTW segmentation is more accurate than HMM, however, the HMM-based phonetic aligners are more reliable. Hence, an hybrid system is proposed in [5]. The results with a Portuguese voice data suggest that the use of HMM-based along with

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 1–11, 2016.
DOI: 10.1007/978-3-319-41552-9_38