Faster time-aligned phonetic transcriptions through partial automation

Ben Serridge and Luciana Castro
Laboratory of Acoustic Phonetics, Universidade Federal do Rio de Janeiro, Brazil

https://doi.org/10.36505/ExLing-2008/02/0050/000109

Abstract

A semi-automatic process for generating time-aligned transcriptions of speech data at the word and phone level is described. At each stage in the process, segment durations are estimated to generate approximate boundary markers, which are then corrected by hand. Corrections at one level are taken into account in the generation of boundaries for the next level, such that the error is reduced at each successive stage. A test implementation based on Praat was applied to a corpus of Brazilian Portuguese and a comparison against a fully manual process revealed a reduction of 54% in the time required to generate phonetic transcriptions and an average error of 21 ms in the time-alignment of phonetic boundaries.

Key words: Brazilian Portuguese, Praat, phonetic transcription, automation tools

Introduction

Linguistics research often relies on access to speech data that has been annotated with time-aligned orthographic and/or phonetic labels. Such corpora are available for heavily studied languages such as English and French, but for most of the world's languages – including major languages such as Brazilian Portuguese – there is very little data available and linguists generally record, transcribe and label their own data as part of their research. The manual transcription process imposes a severe restriction on the amount of data used in the study, and researchers are often forced by time and budget constraints to compromise the robustness of their results.

The ideal solution to the problem is to use an automatic speech recognition (ASR) system configured for forced-path alignment to generate a time-aligned phonetic transcription given the (non-time-aligned) orthographic transcription and a set of grapheme-to-phone rules. Unfortunately, however, there are many languages for which no such ASR system exists, and even for supported languages, the cost, time and technical expertise required to install, configure and successfully apply existing frameworks is prohibitive for many linguistics researchers.

This paper presents a relatively simple process for reducing the time required to transcribe speech data. At each stage in the process, segment durations are estimated using relatively naïve heuristics to generate approximate boundary markers, which are then corrected by hand. Corrections at one level are taken into account in the generation of

boundaries for the next level, such that the error is reduced at each successive stage.

Transcription framework

The framework described in this paper is based on Praat (Boersma 2008), a commonly available speech analysis framework, and leverages the concept of a transcription tier, which enables several layers of time-aligned linguistic annotations for a single utterance, as shown in Figure 1.

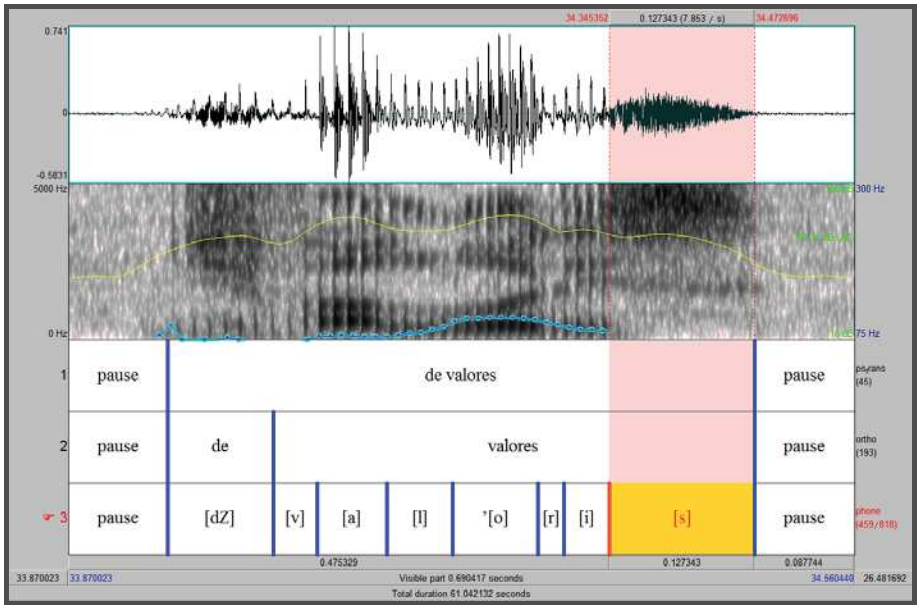

Figure 1. The three transcriptions tiers – phonetic sequence, word-level, and segment-level time-aligned transcriptions – as shown in Praat.

The starting point is a text file in which each line in the file represents the orthographic transcription of a phonetic sequence and line-breaks represent pauses. Taking this file as input, together with the underlying wave file, a Praat script generates the orthographic transcription tier, consisting of alternating intervals representing phonetic sequences (labelled with the transcription itself) and pauses. No pause detection is actually performed on the underlying audio signal; rather, the duration of each pause is fixed (for example, at 500 ms) and the duration of each phonetic sequence is proportional to the number of letters in the orthographic transcription of the sequence, constrained by the overall duration of the phonetic sequences.

Once the phonetic sequence tier has been adjusted by hand, a second script applies a similar procedure to generate a word-level transcription tier,

in which the duration of each word is estimated by multiplying the fraction of letters that the word occupies in the phonetic sequence by the duration of the phonetic sequence as a whole, as given by the previous tier.

Note that up to this point the procedure is fairly language independent. The generation of the phonetic labels, however, requires language-specific rules to predict the set of phones associated with a given orthographic transcription. In this study, the grapheme-to-phone rules described by Silva et al. (2006) were implemented through the use of regular expressions, divided into seven stages, as described in Table 1.

Table 1. Stages in the translation of the orthographic transcription of a word to its phonetic representation (SAMPA).

Phase	Description	Example Transformation
Dictionary Lookup	Handle words whose transcription cannot be predicted by rule.	t á x i → [t] '[a] [k] [s] [i]
Stress Prediction Rules	Mark the vowels that carry primary stress.	c e d o → c 'e d o
Canonical Spelling	Replace known multi-letter combinations with equivalent, unambiguous graphemes.	g 'e s s o → j 'e ç o
Context-dependent Rules	Rules in which the context of the letter determines its mapping.	r a p 'a z → [R] a p 'a [s]
Context-Independent Rules	One-to-one mappings of letters to phones.	[R] a p 'a [s] → [R] [a] [p] '[a] [s]
Standard Phonological Rules	In contrast to the rules applied so far, these operate not on letters but on phones.	[k] '[a] [n] [t] [a] → [k] '[a~] [t] [a]
Regional Phonological Rules	These rules may be applied or not depending on the regional accent of the speaker.	[R] [a] [p] '[a] [s] → [R] [a] [p] '[ai] [S]

The duration of each phonetic unit was calculated based on its average phone duration (Barbosa 1995), scaled appropriately to match the duration of the word as given by the previous tier.

Results

In order to quantify the efficiency gained by applying the above procedure, two one-minute recordings of Brazilian television news were transcribed using Praat: one recording was transcribed following the procedure described in this paper, and the other without the aid of partial automation.

Table 2. The time required to complete each transcription task, with and without the aid of partial automation, expressed in minutes of transcription time per minute of audio, and the average error (in ms) of the predicted boundaries as compared to the final, hand-adjusted boundaries.

Task	Manual Transcription	Partial Automation	% Reduction	Average Boundary Error (ms)
Alignment of Pauses	18	17	6%	672
Word-Level Alignment	50	27	45%	99
Phonetic Transcription	150	55	63%	21
Overall Transcription Task	218	100	54%	N/A

Conclusion

The tools and technical know-how required for fully automating linguistic transcription tasks are inaccessible to most linguists and for the vast majority of the world's languages. Partial automation, however, through the procedures described in this paper, can reduce overall transcription time by half, allowing linguists to work with larger corpora or to spend more of their time on analysis and less on the manual tasks involved in transcription.

References

Barbosa, P. 1995. Estrutura rítmica da frase revelada por aspectos de produção e percepção de fala. Manuscript of talk given at the XLIII Seminário do GEL, May 25-27, 1995, São Paulo, Brazil.

Boersma, P. and Weenink, D. 2008. Praat: doing phonetics by computer (version 5.0.21, http://www.praat.org/).

Silva, D.C., de Lima, A.A., Maia, R., Braga, D., de Moraes, J.F., de Moraes, J.A., and Resende, F.G.V. 2006. A rule-based grapheme-phone converter and stress determination for Brazilian Portuguese natural language processing. Proc. VIth ITS, September 3-6, 2006, Fortaleza, Brazil.