An automatic speech segmentation tool based on multiple acoustic parameters

Introduction

The problem of identification of phrase and utterance boundaries in continuous speech is not a new one. Speech segmentation is required not only for linguistic research based on oral corpora, but had also became essential for natural language processing, such as speech recognition and text-to-speech synthesis, and many researchers have developed different approaches to deal with the need of automatic segmentation of speech data. In this poster we discuss some of these alternatives and present an ongoing project for the automatic segmentation of spontaneous speech developed for Brazilian Portuguese.

Theoretical framework

Several studies on different languages have been demonstrating that prosodic parsing of speech is a highly prominent perceptual phenomenon (1–5). Listeners can detect not only the presence of prosodic boundaries but also differentiate non-terminal from terminal boundaries as well as weak from strong boundaries. Studies with Mandarin show evidence that boundaries are signaled by contrastive neighborhood prosodic states, and that such contrasts are primarily constituted by discourse constraints (6). In this project we adopt the assumption that prosodic boundaries signal the segmentation of spontaneous speech into tone units and utterances (7,8). The term “utterance” is defined here as every linguistic unit that has both pragmatic and prosodic autonomy in discourse, delimited within the speech flow by a terminal prosodic boundary. Utterances can be produced in a single tone unit or they can be parsed into two or more tone units by means of continuative (or non-terminal) prosodic boundaries (8–10). Accurate measures of the acoustic correlates of terminal and non-terminal boundaries are crucial to perform the segmentation of speech into utterances and tone units. The acoustic correlates for prosodic boundaries have been studied for some time. The perception of boundaries is dependent on the occurrence of a set of different intonation features, such as a silent pause, lengthening of the pre-boundary syllable, a rise or fall in f0, as well as change in intensity and also the glottalization over the pre-boundary syllables (11–14). Among these, silent pauses and lengthening of the pre-boundary syllable have been regarded as the most important predictors of boundary perception (2,12,13,15–19).

Aims

The main purpose of this poster is to present a project (currently under development) for the automatic identification of utterance and tone unit boundaries for Brazilian Portuguese. We consider that an automatic segmentation system for spontaneous speech should: 1) be able to identify and differentiate final and non-final boundaries with a minimal margin of error; 2) be based on acoustic data only, and not dependent on syntactic parsing or any other level of previous linguistic analysis; 3) require the least possible amount of human annotation for segmentation training.

In addition, we aim to discuss the main issues related to the two seemingly major strategies adopted for automatic boundary detection, namely, silent pauses and segmental lengthening.

Methods

For training and testing, C-ORAL-BRASIL I (20) corpus is used as source for the speech samples. The project comprises two major components: (a) a qualitative study of inter-annotator agreement for boundary perception, in which we intend to understand what are the determining factors that lead to disagreement among annotators; (b) the development of a script for speech segmentation based on acoustic analysis and boundary perception.

The workflow for the speech segmentation script consists in:

1) Preparation of a speech sample of audio files corresponding to 100 words fragments of texts from different speech styles (monologue and dialogue) and different speakers (male and female). Each audio file is prepared through a Praat (21) annotation object with three tiers: interval tier of all

phonetic syllables defined by two consecutive vowel onsets (VV) (22); point tier with points at every phonological word boundary (potential tone unit boundary locations), with annotation of boundary type: non-boundary, non-terminal boundary or terminal boundary; a second point tier (phonological words) with annotation, for each point, of how many annotators signaled that point as a boundary: 0,1, 2 or 3.

2) Adaptation of a Praat script (23) that uses the corresponding audio file and the annotated tiers to generate the following parameters: mean duration of the phonetic syllables, F0 (median, range, maximum and minimum) and spectral emphasis (24). The script extracts these parameters from a window (10 VV syllables) centered at each potential boundary point.

3) With the acoustic parameters values and the inter-annotator agreement on boundary perception, a logistic regression model is used to predict the likelihood of boundary realization from the acoustic parameters in the sample.

Preliminary results

Observations of spontaneous speech corpus (C-ORAL-BRASIL I corpus; C-ORAL-ROM corpora and the Santa Barbara Corpus) show that final boundaries, i.e. boundaries that delimit utterances (prosodically/pragmatically autonomous linguistic units), can be either perceptually strong or weak, and the same is also true for continuative/non-final boundaries. That means that boundary strength (perceptually weak vs strong boundaries) does not necessarily overlap with boundary type (continuative/non-final and final boundaries), specially in spontaneous speech. Silent pause and pre-boundary syllable lengthening have been successful used as cues to automatic segmentation of speech. However, these parameters seem to be better correlates of boundary strength (weak vs strong boundaries), since neither silent pauses nor lengthening have prove to secure the distinction between final and non-final boundaries. Furthermore, a considerate amount of final and non-final boundaries are not accompanied by silent pauses (around 33% of utterance boundaries and 62% of tone unit boundaries in C-ORAL-BRASIL I corpus). Also, a system based on pre-boundary syllable lengthening for recognition of tone unit boundaries requires the manual syllabic segmentation and annotation of a large volume of data, which takes a great amount of time and skilled human resources. For these reasons, we believe that the extraction of multiple acoustic parameters could provide a more complete probabilistic model for automatic boundary identification in spontaneous speech.

References

1. Sanderman AA. Prosodic phrasing: Production, perception, acceptability and comprehension [Internet]. Technische Universiteit Eindhoven; 1996. Available from: http://alexandria.tue.nl/repository/books/461536.pdf

2. Wightman CW. Segmental durations in the vicinity of prosodic phrase boundaries. J Acoust Soc Am [Internet]. 1992 [cited 2015 Apr 10];91(3):1707. Available from: http://scitation.aip.org/content/asa/journal/jasa/91/3/10.1121/1.402450

3. Mo Y, Cole J, Lee E-K. Naïve listeners' prominence and boundary perception. Speech Prosody [Internet]. Campinas; 2008. p. 735–8. Available from: http://aune.lpl.univ-aix.fr/~sprosig/sp2008/papers/id053.pdf

4. Author

5. Danieli M, Moneglia M, Panizza A, Quazza S, Swerts M, Garrido JM. Evaluation of Consensus

on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech “C-ORAL-ROM.” Prococeedings of the 4th LREC Conference [Internet]. Paris: ELRA; 2004. p. 1513–6. Available from: http://www.lrec-conf.org/proceedings/lrec2004/pdf/371.pdf

6. Tseng C-Y, Chang C-H. Pause or no pause? Prosodic phrase boundaries revisited. Tsinghua Sci Technol [Internet]. 2008 Aug [cited 2015 Apr 28];13(4):500–9. Available from: http://www.sciencedirect.com/science/article/pii/S1007021408700804

7. Cresti E, Moneglia M. Informational patterning theory and the corpus-based description of spoken language: The compositionality issue in the topic-comment pattern. In: Moneglia M, Panunzi A, editors. Bootstrapping Information from Corpora in a Cross-Linguistic Perspective [Internet]. Firenze: Firenze University Press; 2010. p. 13–45. Available from: http://digital.casalini.it/9788884535290

8. Cresti E. Corpus di Italiano parlato. Firenze: Accademia della Crusca; 2000.

9. Moneglia M. A note on spoken language corpora: Units of analysis and language sampling strategies. In: Pizzuto E, Bergman B, editors. 3rd ESF Intersign Workshop: Text corpora and tagging [Internet]. Siena: Universität Hamburg; 1999 [cited 2007 Sep 6]. Available from: http://www.sign-lang.uni-hamburg.de/intersign/workshop3/moneglia/moneglia.html

10. Moneglia M. Spoken Corpora and Pragmatics. Rev Bras Linguística Apl [Internet]. 2011;11(2):479–519. Available from: http://www.periodicos.letras.ufmg.br/rbla/arquivos/335.pdf

11. Nespor M, Vogel I. Prosodic phonology. Dordrecht: Foris; 1986.

12. Campbell N. Automatic detection of prosodic boundaries in speech. Speech Commun [Internet]. 1993 Dec [cited 2015 Apr 27];13(3-4):343–54. Available from: http://www.sciencedirect.com/science/article/pii/016763939390033H

13. Carlson R, Hirschberg J, Swerts M. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. Speech Commun [Internet]. 2005 Jul [cited 2015 Apr 28];46(3-4):326–33. Available from: http://www.sciencedirect.com/science/article/pii/S0167639305000932

14. Blaauw E. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. Speech Commun [Internet]. 1994 Sep [cited 2015 Apr 10];14(4):359–75. Available from: http://www.sciencedirect.com/science/article/pii/0167639394900280

15. Mo Y. Duration and intensity as perceptual cues for naïve listeners’ prominence and boundary perception.

16. Author

17. Shriberg E, Stolcke A, Hakkani-Tür D, Tür G. Prosody-based automatic segmentation of speech into sentences and topics. Speech Commun [Internet]. 2000 Sep [cited 2015 Apr 27];32(1-2):127–54. Available from: http://www.sciencedirect.com/science/article/pii/S0167639300000285

18. Buhmann J, Caspers J, van Heuven VJ, Hoekstra H, Martens J-P, Swerts M. Annotation of prominent words, prosodic boundaries and segmental lengthening by no-expert transcribers in the spoken Dutch corpus. Proceedings of LREC 2002. Paris: ELRA; 2002. p. 779–85.

19. Brierley C, Atwell E. An Approach for Detecting Prosodic Phrase Boundaries in Spoken English. Crossroads [Internet]. New York, NY, USA: ACM; 2007;14(1):5:1–5:11. Available from: http://doi.acm.org/10.1145/1349332.1349337

20. Raso T, Mello H, editors. C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG; 2012.

21. Boersma P, Weenink D. Praat: doing phonetics by computer. [Internet]. 2011. Available from: http://www.praat.org/

22. Author.

23. Author.

24. Traunmüller H, Eriksson A. The frequency range of the voice fundamental in the speech of male and female adults [Internet]. Stockholm; 1994. Available from: http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf