# An automatic speech segmentation tool based on multiple acoustic parameters

Maryualê Malvessi Mittmann°, Plínio Almeida Barbosa*
°Universidade do Vale do Itajaí, *Universidade Estadual de Campinas

Speech segmentation is required not only for linguistic research based on oral corpora, but had become essential for natural language processing. Many researchers have developed different approaches to deal with the need of automatic segmentation of speech data. In this paper we discuss some of the prosodic parameters used as cues for boundary identification and present an ongoing project for the automatic segmentation of spontaneous speech developed for Brazilian Portuguese.

**Keywords:** speech segmentation, spontaneous speech, prosodic boundaries, automatic segmentation

## 1. Introduction

Speech corpora are increasingly becoming important resources for different areas, not only in the field of theoretical and applied linguistics but also for the development of technologies such as text to speech/speech to text systems. However, information extraction from speech corpora requires the segmentation of the audio signal into discrete and meaningful linguistic units. The main goal of this paper is to propose the main features of a model for the automatic segmentation of spontaneous speech based on prosodic parameters in Brazilian Portuguese. We also discuss some of the literature concerning this topic.

The definition for the basic segmental unit of speech may vary according to the researcher's interests. They can be either words or linguistic structures smaller or larger than the word. Many tools have been developed for the segmentation of a speech signal into phonetic units smaller than the word, i.e. phones and syllables. Segmentation at the phonetic level is useful for several purposes, in particular for the extraction of parameters such as duration, fundamental frequency (F0) and intensity within each segment. However, that

type of segmentation is not appropriate for information extraction at the semantic or morphosyntactic levels. Segmentation of the speech signal into words is also not ideal, since it also does not allow a proper extraction and interpretation of various types of linguistic information, like those related to scope and hierarchical relationships between the elements of phrases or other linguistic units that are relevant from a communicative point of view.

In this work we adopt the utterance and the tonal unit as the elementary linguistic units into which the speech flow should be segmented. The boundaries delimiting these units are signalled in the speech flow through prosodic parameters. In the following sections, we present the concepts of these units and discuss some of the literature concerning acoustic parameters related to speech segmentation. We also present a proposal for a model for an automatic speech segmentation tool based exclusively on the analysis of acoustic cues obtained from the audio signal.

## 2. Elementary linguistic units for spoken communication

The problem of identification of phrase and utterance boundaries in speech is not a new one. The idea that prosody is an important component of spoken discourse organization is aknowlegded by a great number of scholars. Since the 1970's, studies of the nature of speech phrasing and the relation between segmental and suprassegmental structures have given rise to different approaches (Halliday 1970; Lehiste 1972; Nespor & Vogel 1986; Chafe 1987; Moneglia & Cresti 1997). Although there is some consensus concerning a relation between prosodic parsing and syntactic structure, the precise nature of this relation is still not fully understood, although an absolute isomorphism between the two domains is no longer advocated.

Several studies on different languages have been demonstrating that prosodic parsing of speech is a highly prominent perceptual phenomenon (Batliner *et al.* 1995; Cummins 1998; Moneglia *et al.* 2010). Listeners can detect not only the presence of prosodic boundaries, but are also able to differentiate discourse finality or continuation according to the perception of non-terminal and terminal boundaries. This is true even when the speech fragments are resynthesised in an unintelligible way by means of spectral filters (Swerts *et al.* 1992), or when the listener is not proficient in the language (Carlson *et al.* 2005).

Studies on different languages have also pointed to a strong connection between prosodic parsing and information/discourse structure (Swerts *et al.*

1992; Chafe 1993; Cresti & Moneglia 2010; Izre'el 2005; Kibrik 2012). There is sufficient evidence from corpus-driven and corpus-based research that the segmentation of speech should be based on prosodic criteria. In accordance with these pieces of evidence, we adopt the assumption that prosodic boundaries signal the segmentation of speech into meaningful communicative units of spoken discourse.

These elementary linguistic units of spoken discourse can be defined within the theoretical framework of the the Language Into Act Theory (Cresti 2000; Moneglia & Raso 2014). According to this theory, the speech flow is parsed into utterances and smaller tonal units by means of prosodic boundaries (Crystal 1975) interpreted by the listener as having either a terminal (concluded/autonomous) or a non-terminal (non-concludes/non-autonomous) value. The term *utterance* is defined as every linguistic unit that has both pragmatic and prosodic autonomy in discourse, delimited within the speech flow by a prosodic boundary perceived as terminal. If the unit carries an illocutionary value (Austin 1962), then the unit is pragmatically (communicatively) autonomous.

Utterances can be produced as a single tonal unit or they can be parsed into two or more tonal units by means of non-terminal prosodic boundaries (Moneglia & Cresti 2006). Example (1) shows a sequence of three simple utterances (Figure 1) and Example (2) shows a compound utterance with two tonal units (Figure 2) (observe the single slash after *sai* in example 2).

1

(1) é a terceira // vão lá // foi // (bpubdl03, 50-52)
'it's the third' 'let's go' 'go'

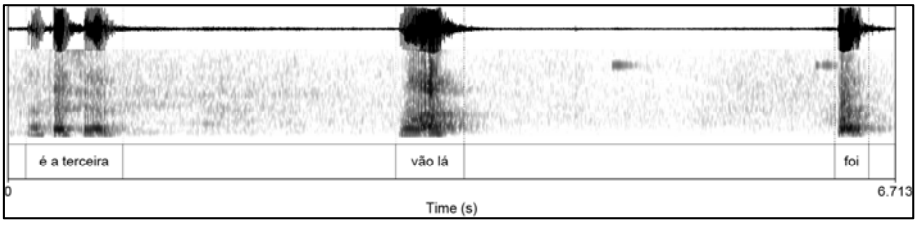


**Figure 1.** Audio wave and spectrogram for example (1).

(2) quando sai / nũ é stop // (bfamdl32, 39)
'when (you're) out' 'it isn't *stop*'

2

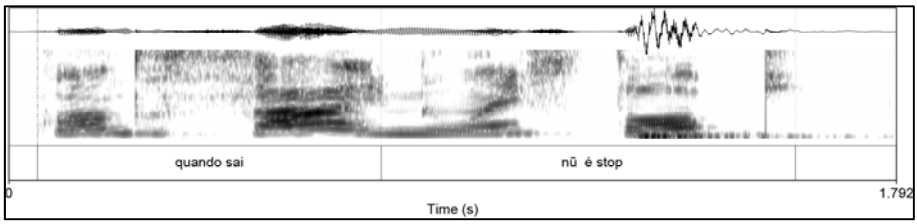


**Figure 2.** Audio wave and spectrogram for example (2)

These examples were selected from the C-ORAL-BRASIL I corpus (Raso & Mello 2012). This corpus comprises 139 informal spontaneous speech recordings and provides audio files, transcriptions and text-to-speech alignment. Double slashes indicate utterance boundaries and single slashes signal tonal unit boundaries. It is important to note that the segmentation of the speech flow into tonal units and utterances is based exclusively on the annotator's perception of terminal and non-terminal prosodic boundaries.

Different acoustic cues appear to be involved in the delimitation of prosodic boundaries. Among these, silent pauses and lengthening of the pre-boundary syllable appear to be the most salient. Examples (3) and (4) show terminal (Figure 3) and non-terminal boundaries (Figure 4) followed by silent pauses.

(3) parece que é colagem / sabe //
'it looks like a colage' 'you know'

olha que interessante // (bfamdl09, 226-227)
'look how interesting'

3

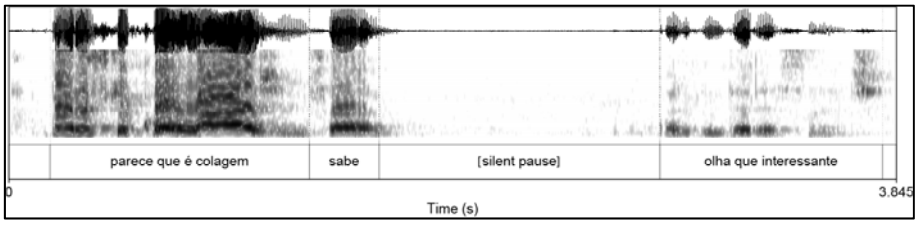


**Figure 3.** Audio wave and spectrogram for example (3)

4

(4) pra poder / manter a casa // (bfammn01, 9)
'so (he) can' 'keep the house'

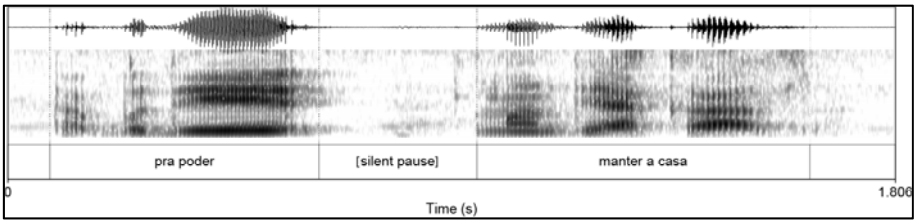


**Figure 4.** Audio wave and spectrogram for example (4)

Examples (5) and (6) show pre-boundary lengthening associated with non-terminal and terminal boundaries respectively. In Figures 5 and 6 the normalized durations[1] of Vowel-to-Vowel (VV) units (Barbosa 2013) were plotted as continuous red lines. Rising lines indicate increase in duration and yellow arrows indicate duration peaks associated with boundaries. VV units are transcribed using a broad phonetic transcription with ASCII characters.

In (5), the stressed vowels on the segments *az* (from *casa* – [k*a*z*a*]) and *ig* (from *barriga* – [baR*i*g*a*]) are both lengthened. These occur in pre- boundary positions. Differently, the non-terminal boundary after the word *não* is not accompanied by pre-syllabic lengthening. This is indicated by the green arrow in Figure 5.

5

(5) eu tô aqui em casa / o Haroldo ainda nũ chegou não /
'I'm here at home' 'Haroldo hasn't arrived yet'

eu tô sentindo assim uma dorzinha na barriga / (bfammn04, 113)
'I'm feeling sorta a little pain in the stomach'

[1] Normalized durations of VV segments are calculated through the z-score (or standard score) method. It consists in extracting the duration of the segment from the audio signal and subtracting it from the mean duration of that particular segment in a given language. The result is then divided by the standard deviation of the duration for the segment in the language. If a VV segment in the audio file has the same duration as the mean duration registered for that segment in the language, its corresponding z-score is zero. Positive z-scores indicate durations higher than average and negative z-scores indicate durations lower that average.

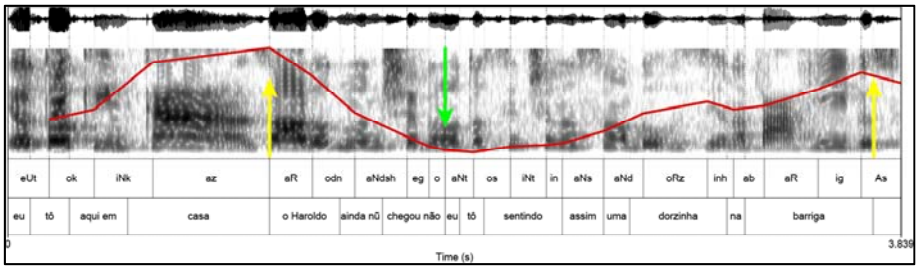


**Figure 5.** Audio wave and spectrogram for example (5) with normalized durations.

In (6), there is a lengthening of the segment *eNtR* (from *ventre* – [v*eNtR*]). This segment is in a pre-boundary position that corresponds to the end of the utterance, therefore, the lengthening here indicates a terminal boundary. This is indicated by the yellow arrow in Figure 6.

(6) você não nasceu do meu ventre // (bfammn05, 63)
'you weren't born of my womb'

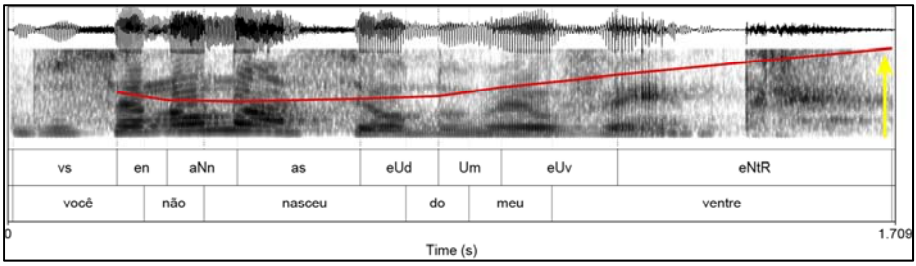


**Figure 6**. Audio wave spectrogram for example (6) with normalized durations.

As can be noted from the examples (3-6), silent pauses and syllabic lengthening are prosodic cues of a boundary. Nevertheless, these parameters alone are not pervasive for the identification of all boundaries. Silent pauses may or may not occur at a perceived boundary. There is an estimate of 33% of utterance boundaries and 62% of tonal unit boundaries that do not coincide with a silent pause in the C-ORAL-BRASIL corpus (Raso *et al.* 2015). Silent pauses and syllabic lengthening are also not good predictors of the boundary type, since they can co-occur with terminal and non-terminal boundaries alike, as exemplified in (3-6).

## 3. Acoustic parameters for the segmentation of speech into discrete units

The acoustic correlates of prosodic boundaries have been studied for some time. The development of powerful and accessible information technologies (personal computer, digital recorders, digital storage media) opened the possibility for the study of naturally ocurring speech fragments using statistical models. From the second half of the 1980's, many studies have been investigating specific aspects of individual as well as groupings of prosodic parameters that signal boundaries.

From these studies, it becomes clear that the perception of boundaries is dependent on the occurrence of a set of different prosodic features, such as a silent pause, lengthening of the pre-boundary syllable, a rise or fall in F0, as well as changes in intensity across the boundary and also creaky voice over the pre-boundary syllables. Among these, silent pauses and lengthening of the pre-boundary syllable have been regarded as the most important predictors of boundary perception and will be further discussed in the following sections.

### 3.1 Silent pause

Silent pause is the most studied parameter of speech segmentation (Martin 1970; Swerts 1997; Shriberg *et al.* 2000; Tseng & Chang 2008; Mo & Cole 2010; Tyler 2013). The analysis of this parameter points at two main results. On the one hand, long pauses are cues related to strong edge marking. On the other hand, the attempt to correlate boundary type (either terminal or non-terminal) with silent pauses shows inconsistent results.

Overall, extra-long pauses are associated with the completion of speech or change of topic (paragraphing). In that case, they are a useful and important parameter for automatic speech processing. However, pauses are not always present at utterance boundaries, and many long pauses do occur between tonal units belonging to the same utterance, as exemplified in (4). Thus, the presence or absence of a silent pause does not provide detailed information about the prosodic parsing of spoken discourse.

### 3.2 Pre-boundary lenghtening

Syllabic lengthening also received a lot of attention on studies on boundary detection. Syllables in pre-boundary position are often much longer than syllables in other positions. (Wightman *et al.* 1992; Barbosa 2008; Mo *et al.* 2008; Fuchs *et al.* 2010; Fon *et al.* 2011; Tyler 2013). This parameter has proved to be quite relevant for automatic speech segmentation.

The syllabic lengthening, however, does not signal speech boundaries only, but can also signal emphasis. Although the portion of the syllable that is lengthened is different in pre-boundary position and when it is used in order to signal emphasis (Campbell 1993; Barbosa 2008), delimiting utterances and tonal units in spoken discourse based on duration would require a very fine analysis (at phone level) for the appropriate training of an automatic segmentation system. Further on, so far there is no evidence that the duration of the pre-boundary syllable could distinguish between two boundary types.

### 3.3 Other parameters

Traditionally, the reset of the fundamental frequency (F0) is considered a cue of intonational phrase boundary. Intonational phrasing can be defined as a structured hierarchy of the intonational constituents in natural speech, dominated by boundary tones (Crystal 2008). The first and last stressed syllables of an intonational phrase delimit a span where there is a gradual declination throughout the whole unit (Couper-Kuhlen 2006). An intonational phrase is usually coextensive with an utterance, but that is not always the case. In fact, some studies have shown that the reset of F0 does not seem to be a sufficient parameter to differentiate boundaries that occur between intonational phrases within an utterance and those delimiting the utterances (Schuetze-Coburn *et al.* 1991; Couper-Kuhlen 2006).

Furthermore, variations in speech rate often signal boundaries between units, as observed in various studies. Generally, a change in the speech rate between the end of an unit and the beginning of the subsequent one is observed (Amir *et al.* 2004; Tyler 2013). What is more, this parameter is closely related to the style of speech and the particular characteristics of the speaker.

Intensity is also used as an auxiliary parameter in boundary identification, since it exhibits a declination line similar to F0 declination. Moreover, an increase in intensity can be related to the beginning of a prosodic unit (Swerts *et al.* 1994; Tseng & Fu 2005; Mo 2008).

Finally, laryngealization (creaky voice) has also been pointed out as an acoustic cue for the identification of prosodic boundaries. Studies on different languages indicate that laryngealization occurs mainly at prosodic boundaries, that is, the final portion of utterances or intonational phrases are often creaky (Kohler 1994; Redi & Shattuck-Hufnagel 2001; Ogden 2001; Garellek 2015) and it seems to be also related to fragmentation and disfluency phenomena (Kohler 1994; Kohler *et al.* 2001).

## 4. A model for an automatic speech segmentation tool

From the literature reviewed above, it is clear that an automatic speech segmentation task will not be accurate if it is based on a single prosodic parameter. Particularly, the task of differentiating between terminal and non-terminal prosodic boundaries cannot be achieved without a better understanding of the possible sets of parameters that correlate with each boundary type. Therefore, accurate measures of the acoustic correlates of terminal and non-terminal boundaries are crucial to perform the segmentation of speech into utterances and tonal units.

We consider that an automatic segmentation system for spontaneous speech should:

- be able to identify and differentiate terminal and non-terminal boundaries with a minimal margin of error;
- be based on acoustic data only, and not dependent on syntactic parsing or any other level of previous linguistic analysis;
- require the least possible amount of human annotation for segmentation training.

To achieve these goals, we adopt the following procedures: first, we prepare a speech sample of audio files. The files correspond to 100-200 words fragments of texts from different speech styles and speakers (male and female). Speech samples are selected from the C-ORAL-BRASIL Reference corpus for spontaneous Brazilian Portuguese (Raso & Mello 2012; Raso & Mello 2014). Excerpts are taken from spontaneous monologues in informal and formal natural contexts and also from media (news). Each audio transcript is then annotated by a team of 14 expert prosodic boundary annotators. Annotators work independently of one another. While they listen to the audio, they insert tags for non-terminal and terminal prosodic boundaries on the transcript according to their perception. All boundary tags are counted according to type (non-terminal and terminal) and position (pre-boundary phonological word). This information is then transferred to point tiers in a Praat TextGrid object[2].

Next, the audio files are annotated with Praat TextGrid objects with five tiers:

[2] http://www.praat.org/ (accessed December 5, 2015).

- Vowel-to-Vowel interval tier. Interval tier with all phonetic syllables delimited by two consecutive vowel onsets accompanied by a broad phonetic transcription;
- Phonological Word point tier – non-terminal boundary. Point tier with points at every phonological word boundary (potential tonal unit boundary locations), accompanied by a label, at each point, of how many annotators signaled that point as a non-terminal boundary: 0-14.
- Phonological Word point tier – terminal boundary. Point tier with points at every phonological word boundary, accompanied by a label, at each point, of how many annotators signaled that point as a terminal boundary: 0-14.
- Silence. Interval tier delimiting silent pauses.
- Text. Textual transcription of utterances.

In order to generate a model for the automatic annotation of prosodic boundaries, the Praat script *ProsodyDescriptor* (Barbosa 2013) is being adapted for the extraction of prosodic parameters at each point indicated in the Praat annotation object. The script uses the corresponding audio file and the annotated tiers to extract and calculate the following parameters:

*Measures for speech rate and rythm:*

a. speech rate in VV units per second;
b. rate of non-salient VV units per second;

*Measures for segment duration and duration normalization[3]:*

c. Mean, standard deviation and skewness of smoothed z-score peaks;
d. Smothed z-scored local peak rate in peaks per second;

*Measures for fundamental frequency (F0) and F0 normalization:*

e. F0 median in semi-tones (re 1 Hz);
f. F0 standard deviation in semi-tones;
g. F0 Pearson skewness[4];
h. 1st-derivative F0 mean in Hz/second – the value is multiplied by 1000 for scaling purposes;
i. 1st-derivative[5] F0 standard deviation in Hz/second;
j. 1st-derivative F0 skewness;

[3] See footnote 1.

[4] Skewness is a statistical measure of symmetry (or lack of it) in a distribution.

[5] The first derivative indicates whether a function is (and by how much) increasing or decreasing. All statistical measures extracted here are used to determine the direction of the pitch movement (rising or falling) and the shape of a smoothed and normalized pitch curve.

k. smoothed F0 peak rate in peaks per second;

*Measure for intensity:*

l. spectral emphasis in dB.

The script extracts these parameters from a window of 10 VV units to the left and 10 VV units to the right of each potential boundary. Figure 7 shows an example of the audio and annotation grid in Praat with the analysis windows (shaded in yellow) for the boundary point indicated by the red arrow.

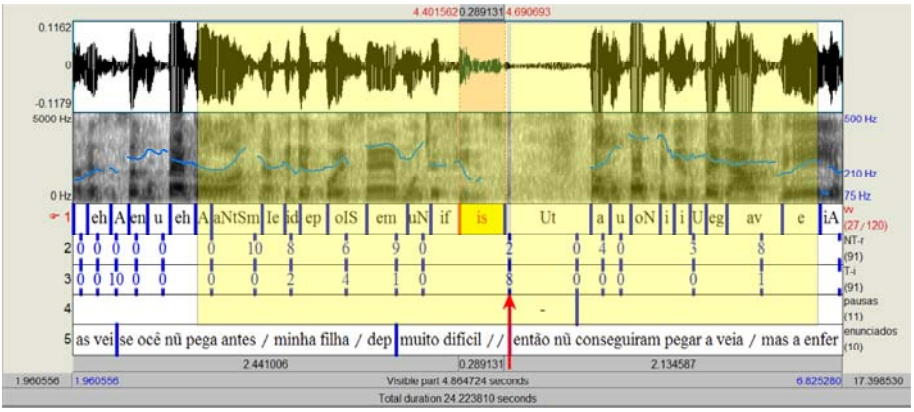


**Figure 7.** Illustration of audio file, Textgrid and analysis windows for data extraction.

With the acoustic parameters values and the inter-annotator agreement on boundary perception, a logistic regression model will be used to predict the likelihood of boundary realization from the acoustic parameters in the sample. And this for the two types of boundary.

## 5. Final remarks

Observations of spontaneous speech corpora such as C-ORAL-BRASIL I (Raso & Mello 2012), C-ORAL-ROM (Cresti & Moneglia 2005) and the Santa Barbara Corpus (Du Bois *et al.* 2000-2005) 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 (terminal and non-terminal boundaries), specially in spontaneous speech.

Silent pause and pre-boundary syllable lengthening have been successfully used as cues to automatic segmentation of speech. However, these parameters seem to be better correlates of boundary strength (perception of weak vs. strong boundaries) than of boundary type (terminal vs. non-terminal), since neither has proved to guarantee the distinction between final and non-final boundaries. Also, a system based on pre-boundary syllable lengthening for recognition of tonal 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.

So far, we do not have a model that correlates different sets of prosodic parameters with terminal and non-terminal boundaries, as proposed in this paper. For the reasons pointed out above, we believe that the extraction of multiple acoustic parameters could provide a more complete probabilistic model for automatic boundary identification in spontaneous speech.

## References

Amir, N., Silber-Varod, V. & Izre'el, S. 2004. Characteristics of intonation unit boundaries in spontaneous spoken Hebrew: Perception and acoustic correlates. In B. Bell & I. Marlien (eds), *Speech Prosody 2004: Proceedings*. Nara: ISCA, 677-680. http://sprosig.isle.illinois.edu/sp2004/PDF/Amir-SiberVarod-Izreel.pdf.

Austin, J.L. 1962. *How to do things with words*. Oxford: Clarendon Press.

Barbosa, P.A. 2008. Prominence- and boundary-related acoustic correlations in Brazilian Portuguese read and spontaneous speech. In P.A. Barbosa, S. Madureira & C. Reis (eds), *Speech Prosody*. Campinas: ISCA, 257–260. http://aune.lpl.univ-aix.fr/~sprosig/sp2008/papers/id060.pdf.

Barbosa, P.A. 2013. Semi-automatic and automatic tools for generating prosodic descriptors for prosody research. In B. Bigi & D. Hirst (eds.), *TRASP 2013 Proceedings*, Vol. 13. Aix-en-Provence: Laboratoire Parole et Langage, 86–89. http://www.lpl-aix.fr/~trasp/Proceedings/19874-trasp2013.pdf (accessed December 22, 2015).

Batliner, A., Kompe, R., Kießling, A., Niemann, H., E. Nöth & Kilian, U. 1995. The Prosodic Marking of Phrase Boundaries: Expectations and Results. In A.J. Rubio Ayuso & J.M. Lopez-Soler (eds.), *Speech Recognition and Coding: New advances and Trends*, vol. 147, 89–92. Berlin: Springer. http://link.springer.com/chapter/10.1007/978-3-642-57745-1_48 (accessed December 22, 2015).

Du Bois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A., Englebretson, R. & Martey, N. 2000-2005. *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium. http://www.linguistics.ucsb.edu/research/santa-barbara-corpus (accessed December 5, 2015).

Campbell, N. 1993. Automatic detection of prosodic boundaries in speech. *Speech Communication* 13(3-4): 343–354. http://www.sciencedirect.com/science/article/pii/016763939390033H (accessed April 27, 2015).

Carlson, R., Hirschberg, J. & Swerts, M. 2005. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication* 46(3-4): 326–333. http://www.sciencedirect.com/science/article/pii/S0167639305000932 (accessed April 28, 2015).

Chafe, W.L. 1987. Cognitive Constraints on Information Flow. In R.S. Tomlin (ed.), *Coherence and Grounding in Discourse*. Amsterdam: Benjamins.

Chafe, W.L. 1993. Prosodic and Functional Units of Language. In J.A. Edward & M.D. Lambert (eds), *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Lawrence Erlbaum Associates, 33-43

Couper-Kuhlen, E. 2006. Prosodic Cues of Discourse Units. In K. Brown (ed.), *Encyclopedia of Language & Linguistics*, 2nd ed. Amsterdam: Elsevier, 178–182.

Cresti, E. 2000. *Corpus di Italiano parlato*, Vol. 1. Firenze: Accademia della Crusca.

Cresti, E. & Moneglia, M. (eds) 2005. *C-ORAL-ROM: Integrated reference corpora for spoken romance languages*. Amsterdam: John Benjamins.

Cresti, E. & Moneglia, M. 2010. Informational patterning theory and the corpus-based description of spoken language: The compositionality issue in the topic-comment pattern. In M. Moneglia & A. Panunzi (eds), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*, Firenze: Firenze University Press, 13–45.

Crystal, D. 1975. Intonation and Linguistic Theory. In K.H. Dahlstedt (ed.), *The Nordic Languages and Modern Linguistics 2: Proceedings of the Second International Conference of Nordic and General Linguistics*. Stockholm: Almqvist & Wiksell, 267–303.

Crystal, D. 2008. *A dictionary of linguistics and phonetics*, 6th ed. Oxford: Blackwell.

Cummins, F. 1998. Rhythmic constraints on stress timining in English. *Journal of Phonetics* 26: 145–171.

Fon, J., Johnson, K. & Chen, S. 2011. Durational patterning at syntactic and discourse boundaries in Mandarin spontaneous speech. *Language and speech* 54(1): 5–32.

Fuchs, S., Krivokapić, J. & Jannedy, S. 2010. Prosodic boundaries in German: Final lengthening in spontaneous speech. *The Journal of the Acoustical Society of America* 127(3): 1851.

Garellek, M. 2015. Perception of glottalization and phrase-final creak. *The Journal of the Acoustical Society of America* 137(2): 822–831.

Halliday, M.A.K. 1970. Language structure and language function. In J.J. Webster (ed.), *On grammar*. London: Continuum, 173–195.

Izre'el, S. 2005. Intonation Units and the Structure of Spontaneous Spoken Language: A View from Hebrew. Paper presented at the *International Symposium on "Towards modeling the relations between prosody and spontaneous spoken discourse"*, Aix-en-Provence, France.
https://pdfs.semanticscholar.org/b561/ec35cda51091e021fea25d411025092f119b.pdf?_ga=1.35796327.633103945.1481736294 (

Kibrik, A.A. 2012. Prosody and local discourse structure in a polysynthetic language. In Y. I. Alexandrov (ed.), *Fifth International Conference on Cognitive Science*, Vol. 1. Kaliningrad: MAKI, 80–81.

Kohler, K.J. 1994. Glottal Stops and Glottalization in German. *Phonetica* 51(1-3): 38–51.

Kohler, K.J., Peters, B. & Wesener, T. 2001. Interruption Glottalization in German Spontaneous Speech. In *Disfluency in Spontaneous Speech (Diss01)*, 45–48. http://www.isca-speech.org/archive_open/archive_papers/diss_01/dis1_045.pdf (accessed December 1, 2016).

Lehiste, I. 1972. The Timing of Utterances and Linguistic Boundaries. *The Journal of the Acoustical Society of America* 51(6): 2018-2024.

Martin, J.G. 1970. On judging pauses in spontaneous speech. *Journal of Verbal Learning and Verbal Behavior* 9(1): 75–78.

Mo, Y. 2008. Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception. In P. A. Barbosa, S. Madureira & C. Reis (eds), *Speech Prosody*, Campinas: ISCA, 739–742.

Mo, Y. & Cole, J. 2010. Perception of prosodic boundaries in spontaneous speech with and without silent pauses. *The Journal of the Acoustical Society of America* 127(3): 1956.

Mo, Y., Cole, J. & Lee, E.K. 2008. Naïve listeners' prominence and boundary perception. In P.A. Barbosa, S. Madureira & C. Reis (eds), *Speech Prosody*, Campinas: ISCA, 735–738.

Moneglia, M. & Cresti, E. 1997. L'intonazione e i criteri di trascrizione del parlato adulto e infantile. In U. Bortolini & E. Pizzuto (eds), *Il Progetto CHILDES Italia*. Pisa: Del Cerro, 57–90.

Moneglia, M. & Cresti, E. 2006. C-ORAL-ROM. Prosodic boundaries for spontaneous speech analysis. In Y. Kawaguchi, S. Zaima & T. Takagaki (eds), *Spoken Language Corpus and Linguistics Informatics*. Amsterdam: Benjamins, 89–112.

Moneglia, M. & Raso, T. 2014. Notes on Language into Act Theory (L-AcT). In T. Raso & H.R. Mello (eds), *Spoken Corpora and Linguistic Studies*, Amsterdam/Philadelphia: John Benjamins, 468–494.

Moneglia, M., Raso, T., Mittmann, M.M. & Mello, H.R. 2010. Challenging the Perceptual Relevance of Prosodic Breaks in Multilingual Spontaneous Speech Corpora: C-ORAL-BRASIL / C-ORAL-ROM. In *Speech Prosody 2010* 102010:1-4. http://speechprosody2010.illinois.edu/papers/102010.pdf (accessed December 1, 2016).

Nespor, M. & Vogel, I. 1986. *Prosodic phonology*. Dordrecht: Foris.

Ogden, R. 2001. Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association* 31(01): 822-831.

Raso, T. & Mello, H. 2014. C-ORAL-BRASIL: Description, Methodology and Theoretical Framework. In T. Berber Sardinha & T. L. São Bento (eds), *Working with Portuguese Corpora*. London: Bloomsbury, 257–278.

Raso, T., Mittmann, M.M. & Mendes, A.C.O. 2015. O papel da pausa na segmentação prosódica de corpora de fala. *Revista de Estudos da Linguagem* 23(3): 883–922.

Raso, T. & Mello, H. (eds). 2012. *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: UFMG.

Redi, L. & Shattuck-Hufnagel, S. 2001. Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* 29(4): 407–429.

Schuetze-Coburn, S., Shapley, M. & Weber, E.G. 1991. Units of intonation in discourse: a comparison of acoustic and auditory analyses. *Language and speech* 34(3): 207–234.

Shriberg, E., Stolcke, A., Hakkani-Tür, D. & Tür, G. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* 32(1-2): 127–154.

Swerts, M. 1997. Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America* 101(1): 514–521.

Swerts, M., Collier, R. & Terken, J. 1994. Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication* 15(1-2): 79–90.

Swerts, M., Geluykens, R. & Terken, J. 1992. Prosodic correlates of discourse units in spontaneous speech. In J.J. Ohala (ed.), *Proceedings of the International Conference on Spoken Language Processing*, Banff: ISCA, 421–424.

Tseng, C.Y. & Chang, C.H. 2008. Pause or no pause? Prosodic phrase boundaries revisited. *Tsinghua Science and Technology* 13(4): 500–509.

Tseng, C.Y. & Fu, B. 2005. Duration , Intensity and Pause Predictions in Relation to Prosody Organization. *Interspeech 2005*, 1405–1408. http://www.ling.sinica.edu.tw/eip/FILES/publish/2007.4.12.99500673.0143164.pdf (accessed December 1, 2016).

Tyler, J. 2013. Prosodic correlates of discourse boundaries and hierarchy in discourse production. *Lingua* 133: 101–126.

Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P.J. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America* 91(3): 1707–1717.

## Acknowledgments

The authors are grateful to FAPEMIG for financing this work.