Barbosa, P. A. (no prelo) *Proc. of the GSCP 2012 Workshop*, Belo Horizonte 29/2 a 2/03. Firenze University Press.

Panorama of Experimental Prosody Research

Plínio A. Barbosa
Instituto de Estudos da Linguagem, State University of Campinas, Brazil
Rua Sérgio Buarque de Holanda, 571 Zip code: 13083-859 - Campinas-SP, Brazil
E-mail: pabarbosa.unicampbr@gmail.com

Abstract

Starting by a definition, this paper presents a panorama of experimental prosody research. After briefly exposing the three main properties of an experimental work, testability, predictability and designability, as well as the selection of variables in experimental prosody research, the crucial concepts and methodological procedures involved in rhythm and intonation research are portrayed. These aspects are presented in functional terms, that is, they are explored as a means to reveal the functions of prosody in verbal communication. The concepts of prominence, acoustic salience, pitch accent, prosodic boundary, stress group, phrase stress, speech rate, phonetic syllable, pausing, tonal alignment and expressive speech are presented and illustrated with some examples from work on Brazilian Portuguese and Standard German. The procedures depicted in this work were the duration normalisation technique for rhythm research and the pitch accent and boundary tone annotation in intonation research. Some questions such as the terminological difference between "intonation" and "prosody", the stress- vs syllable-timing distinction, the difference between perceived and produced prosody are briefly presented and discussed.

Keywords: Prosody, Experimentation, Rhythm, Intonation, Expressive Speech

1. Introduction

Experimental prosody can be defined as the area of research which applies the hypothetic-deductive method to prosodic studies via experimentation. This definition implies that experimentation in prosody research should preferably be developed in three steps of increasing complexity: observation, description and experimentation *stricto sensu.*

The observation of a prosodic fact is never naive, because formal instruction is necessary to see or to select what is relevant in terms of a variable under scrutiny (for a general reading about observation in science see Fleck, 1992, Beveridge, 1957, p. 102-105 and Bunge, 1998, p. 181-189). As an illustration, fundamental frequency (F_0) peaks can be of different heights but only some are relevant from the perceptual of from the linguistic points of view. Thus, a simple question such as "what is a linguistically meaningful F_0 peak?", needs a formal instruction to be appropriately answered.

Descriptive prosodic research is an important step of scientific discovery. It uses the formal devices of descriptive statistics or correlational methods to give measures of centrality, variation, amplitude and skewness in the former case or the correlation between two or more variables in the latter case. Several other measures can be used; we presented here the most common ones. The statistical descriptors reduce the degrees of freedom of the variables and give a first picture of the phenomena under scrutiny.

Experimentation is related to reproducibility, which is a key scientific component. That is why this step is so closely related to inferential statistics: "One of the first things which the beginner must grasp is that statistics need to be taken into account when the experiment is being planned, or else the results may not be worth treating statistically." (Beveridge, 1957, p. 19). Under certain conditions of control, a snapshot of a communicative instance (the corpus) is examined and the variables of interest are measured to infer, given the variation of the data, the behaviour of a population from which the data were obtained. Experimentation starts with a theory, which guides the observation of prosodic facts. The theory and the observed facts produce a set of hypotheses aiming at testing a model of prosody production or perception. To test this model, a set of hypothesis-derived measures extracted from the corpus are evaluated according to their validity as regards the hypotheses raised at the beginning of the experimental study. This last step allows the refinement or revision of the theory that motivated the study.

In section 2 some considerations and initial steps for carrying out an experimental work are given. In section 3, we present the functions of prosody. In sections 4 and 5 we respectively present the main conceptual and methodological in rhythm and intonation research. Section 6 gives some directions and key concepts of expressive speech research. The aim of this paper is not to present a review of prosodic research, but to give a panorama of experimental prosody research to stimulate the new comer to choose an area of research to work with.

2. Getting started in experimentation

2.1. Properties of an experimental work

In order to be scientific valid, a theory in experimental prosody research needs to satisfy three main properties: testability, predictability, and designability (for a similar view, see Xu, 2011).

Testability refers to the hypotheses raised by the experimenter. They should be well-formed, meaningful, and contain mechanisms to check whether they are true or false (Bunge, 1998, p. 309-315). The truth-conditions of an original hypothesis can be refined after experimentation, but the reformulated hypothesis should also be directly testable. Suppose that an experimenter

posits the hypothesis that stressed syllables are longer than unstressed syllables in Brazilian Portuguese (henceforth BP) based on previous experimental findings that suggested that syllable duration is the main correlate of stress in BP (Martini, 1991; Barbosa, 1996). This hypothesis is testable because we can design a corpus for comparing the duration of stressed syllables with that of unstressed syllables in similar conditions of production and then apply a two-sample statistical test to compute the probability of making a type-I error when rejecting the null hypothesis that both durations are the same. This does not mean that this kind of check is easy, considering the several elements that affect duration needing to be controlled. For instance, at the end of an utterance, an unstressed syllable with an identical phonemic composition of a stressed syllable (e.g., the second syllable of *papa* – pope) is longer than the latter because of the final lengthening phenomenon (Scott, 1980; Edwards et al., 1991). This simple exception to the general finding would entail the refinement of the original hypothesis to: "non-pre-pausal unstressed syllables are shorter than stressed syllables". Additional exceptions can be discovered from subsequent experimental settings.

Predictability refers to the ability of predicting new outcomes under distinct experimental conditions. In order to do so, how to predict the values of the new outcomes must be explicit. This explicitness is associated with a model which can be conceived of as a set of rules or a set of equations. For instance, intonation models and rhythm models can generate F_0 and duration values for a particular utterance, which can be compared with the observed utterance for a certain number of speakers to assess the closeness between predicted and observed values, and, given the nature and/or extension of the errors obtained, evaluate the need for model refinement (some examples of either rhythm or intonation models can be found in Xu, 2011; Botinis et al., 2001; Barbosa, 2006).

Designability refers to the possibility of conceiving an experimental protocol to test the hypotheses raised. The design of an experiment in prosody research is not easy. It includes the selection of variables for investigation, the choice of the statistical test to assess the hypotheses, the choice of the informants to record the corpus or of the subjects to listen to the set of stimuli of a perception test. The example of the stressed vs unstressed syllable mentioned before presents a high degree of designability. But it is not always like that. Suppose that a theoretical account of the relation between neuronal activity and speech perception states that a particular pathway is more activated when a subject listens to a C-to-V transition. Two ideal experimental designs could be: (1) to put electrodes directly in the areas along the pathway and to measure neuronal activity or (2) to make a lesion in some area in the pathway and study its consequences. It is unnecessary to explain the ethical problems involved in both designs. Researchers can cope with them by studying the aforementioned relationship in non-human mammals or by studying the consequences of naturally-occurring lesions in human patients (see some studies reported by Scott & Wise, 2003).

2.2. The selection of the variables for study

There are three classes of variables in an experimental setting: independent, dependent and to-be-controlled. Independent variables are those manipulated by the experimenter and directly related to the hypotheses raised. It is important to know that they are not necessarily nominal or discrete.

Dependent variables are those which are measured and which are usually acoustic or articulatory correlates of the discrete or intervalar prosodic, independent variables.

To-be-controlled or nuisible variables are those that need to be controlled because their unpredicted (or unpredictable) influence can affect the dependent variables if we do not take enough care.

Let's examine three examples of these variables in prosodic research. First, the experiment about the duration of stressed vs unstressed syllables in BP presented above. In this case, the independent variable is STRESS, with two levels, "stressed" and "unstressed". The dependent variables are the acoustic duration of the syllables. The to-be-controlled variables are: phonetic context of the syllable, degree of prominence and boundary adjacent to the measured syllables, speech rate, and healthy state of the subject, among others. We cannot compare stressed vs unstressed syllables in words where the to-be-controlled variables differ because the non-chosen differences in these variables can also affect duration. In this case, it is not possible to infer the cause of the duration change. For instance, a stressed syllable in a word just after a previous focussed word can exhibit lesser duration than an unstressed syllable with a similar phonetic context in a word not in post-focal position. The ideal statistical test for comparing the mean duration of the syllables across the two levels of the STRESS variable is a t-test of independent variables, provided that the residue is normally distributed (otherwise the equivalent non-parametric test is Mann-Whitney. See Crawley, 2005 for a nice introduction and use of statistical tests).

As a second example, suppose you want to determine how many distinct boundary levels can be signalled by a relevant acoustic-prosodic parameter in BP. The independent variable is the height of the constituent immediately preceding the boundary in a hierarchy of linguistic domains. The dependent variable can be the duration of the syllable rhyme preceding the boundary (cf Barbosa, 2006). The to-be-controlled variables are all extraneous variables that affect the duration of the pre-boundary words but the boundary height in the hierarchy. The appropriate statistical test is clusterisation, which groups together, under certain conditions, the durations associated to the same statistical distributions. At the end of the process, the number of distinct boundary levels is the number of distinct statistical groups (see Whitman et al., 1992 for research of boundary levels in American English, Barbosa, 1994 for Standard French and Barbosa, 2006 for BP).

The final example concerns expressive speech. Suppose you want to predict the degree of arousal evaluated by a group of listeners from the acoustic-prosodic properties of an utterance. In this case, the independent variable is the set of acoustic-prosodic parameter values for the utterance. The dependent or predicted variables are the listeners' evaluation degrees, and the appropriate statistical test is multiple regression. The to-be-controlled variables are all the influencing factors that could explain the listeners' evaluation which are not based on what they hear from the acoustic information embedded in the speech signal, such as the lexicon, the habit of a listener in giving high grades, the health state of the listener that day, among others.

3. Functions of prosody

In terms of linguistic and paralinguistic uses, the following functions of prosody can be identified: (1) a discursive function such as to signal a turn in a dialogue, to signal that you are listening your interlocutor (backchannels such as "um-hum", "entendo" – (I) understand), to signal the modality of a sentence in a monologue, (2) a demarcative function aiming at signalling the edges of a prosodic constituent such as a phonological word or a stress group, (3) a prominence function aiming at signalling to the listeners the salience of a prosodic unit in relation to another one or in relation to the background units (see Barbosa, submitted, for examples of these functions and an introduction to prosodic research).

In terms of expressiveness, the following functions can be distinguished: attitudinal (attitude, personal stance) and affective (emotions such as sadness, joy and rage as well as other affects such as humour and traits of personality). Prosody can also convey indexical features such as gender, sex, dialectal and social origin, among others. Expressive and indexical features are found in every single utterance produced by a human subject because it's very hard to disguise aspects such as attitude, emotion and sex.

For an introduction to expressive speech research see the works by Fónagy (1986), Bolinger (1986) and Scherer (1984).

4. Conceptual and methodological aspects in rhythm research

More than a hundred definitions of rhythm can be given. Several of those proposed by Sauvanet (2000) highlight, in my sense, the two main components of rhythm, structuring and repetition: "Il y a rythme lorsqu'une structure évolue de manière périodique sur fond d'altération novatrice." (Wunenburger, 1992, p. 17) and "The essence of rhythm is the fusion of sameness and novelty; so that the whole never loses the essential unity of the pattern, while the parts exhibit the contrast arising from the novelty of their detail." (Whitehead, 1919, p. 198). Thus, there is periodicity and structuring in speech rhythm. Periodicity (sameness in Whitehead's terms) serves the production system because it makes the control of the units produced easier. But an utterance with identical units would never signal anything to the listener. Then, it's necessary to build a structure to differ (novelty in Whitehead and Wunenburger) from the background. But what is repeated and what is modified to signal novelty? Essentially, syllables.

In BP, when a word is produced with acoustic salience, the acoustic parameters around the lexically stressed syllable are modified in relation to the background formed by the non-salient syllables. These acoustic parameters are F_0, duration, intensity, formant values, among others. In BP, salient syllables are often longer than non-salient ones. At strong syntactic boundaries or to signal a focussed item, these syllables are also higher in pitch (Barbosa, 2008). If the acoustically salient syllable is audible we say that the syllable and the word containing it are prominent, because these units catch the attention of the listener. Rhythm is the sensation caused by the succession of different degrees of syllabic prominence alternated with non-prominent syllables in the background (Barbosa, 1994).

Nowadays, rhythm research deals with the study of patterns of syllable-size duration along the utterances. In order to do so, it's necessary to separate segmental from prosodic information of syllable-sized durations. This is done by a technique of normalisation.

Duration normalisation allows to highlight with an accuracy of up to 80 % (Barbosa, 2010), the phonological words perceived as prominent or pre-boundary by the listeners. This is done by detecting normalised syllable-sized durations peaks in three steps. In the first step, the z-score of the phonetic syllable duration is computed. The phonetic syllable starts at the vowel onset of the realised phonological syllable and ends at the vowel onset of the next realised phonological syllable and is known in the literature as V-to-V unit (Barbosa, 2006). It has been used in rhythm research since a long time (cf Lehiste, 1970; Classe, 1939). By definition, the z-score, a common statistical measure, expresses the distance from the mean in units of standard-deviation. Then, if a z-score is -1.3 (it has no physical unit), this means that the duration is 1.3 standard-deviations distant from the mean, leftwards. Mean and standard-deviation can be obtained from a corpus containing all phones of a language, and it does not need to be from the same speaker, although it is recommended that the subject be from the same dialectal area (cf. Barbosa, 2006, p. 489 for values for these two descriptors in BP).

In the second step, a 5-point moving average technique is used to filter out additional sources of variation not related to perceived duration (for mathematical details see Barbosa, 2010 and Barbosa, 2006). The normalisation aims at minimising the effects of intrinsic duration and those of the number of segments of the V-to-V units.

The result of these two steps can be seen in Fig. 1 for the utterance "Manuel tinha entrado para o mosteiro há quase um ano, mas ainda não se adaptara àquela maneira de viver.", uttered by a female speaker from São Paulo State. In the figure, five duration peaks around the respective stressed syllables of five words can be seen:

"entrado", "mosteiro", "ano", "adaptara", "viver". The three higher peaks are indicated, perceived by all listeners as pre-boundary or prominent. The peaks within the two other words are perceived as weakly prominent words. The peaks of normalised duration can be automatically detected by tracking the points where the derivative of the contour changes from positive to negative, which is the third step.

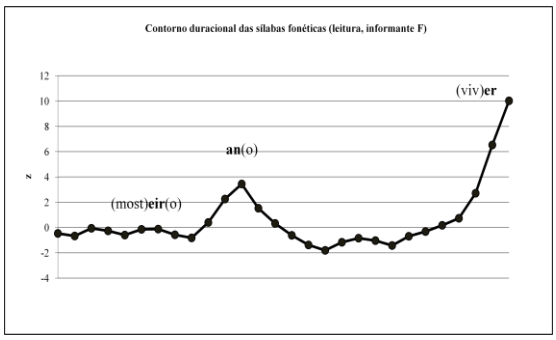

Figure 1: V-to-V normalised duration of the utterance "Manuel tinha entrado para o mosteiro há quase um ano, mas ainda não se adaptara àquela maneira de viver." by a female speaker.

Each normalised V-to-V duration peak indicates an acoustic salience that, if perceived as a prominence, represents the position of a phrase stress. Because in BP the probability of a pre-boundary word be perceived as prominent is between 40 and 65 % (Barbosa, 2008) and because boundaries define the end of a domain, the association of normalised duration peaks to stress group boundaries is a convenient and appropriate decision for the need of automatic stress group delimitation. Despite the signalling of both prominence and prosodic boundary by longer durations, it is possible to distinguish the two functions when looking at the consequences of their implementation for the segments that make up the syllables. This difference can be found at least if the speaker highlights a word for signalling emphasis. In emphatic words, all segments of the lexically stressed syllable are lengthened, whereas, for words before a prosodic boundary, the stressed phonetic syllable is lengthened, that is, the vowel and the consonants following it (tautosyllabic or heterosyllabic). As an example in BP, let's choose the two sentences "Pedro vai casar, sabia?" and "Pedro vai CASAR, sabia?" The segments /a/ and /R/ are much more lengthened than /z/ in the word "casar" in the first sentence, whereas the segments /z/, /a/ and /R/ of the entire stressed syllable of the emphatic word in the second sentence are equally lengthened. This prosodic fact was experimentally demonstrated by Barbosa (2006, p. 309-317), and is found in several languages (see Tabain, 2003 for French and Byrd and Saltzman, 1998 for American English).

Another striking result of the normalisation technique is that the height of the duration peaks closely follows the degree of strength of the prosodic boundaries or prominences: the strongest boundary is after the word "viver", followed by that after the word "ano". Without the application of this technique, the raw duration peak position and height do not correspond to valid prosodic functions as can be seen in Fig.2, where there are 12 peaks of duration. No listener perceives 12 prominent or pre-boundary words in this utterance. The normalisation procedure is basic in rhythm research and should be followed to reveal prosodic duration.

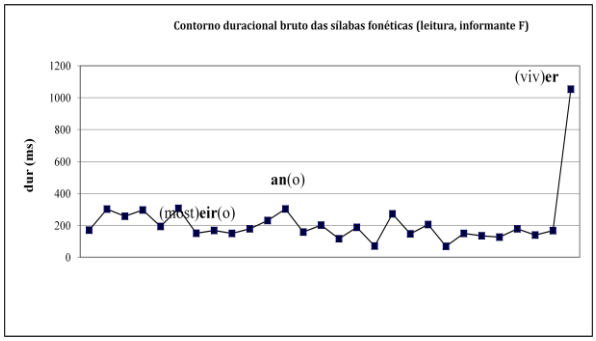

Figure 2: V-to-V raw duration contour of the same utterance of Fig. 1

By analysing V-to-V normalised duration and other acoustic parameters such as vowel formant values, F_0 and spectral emphasis (Traunmüller & Eriksson, 2000), Arantes (2010) showed that duration and F_0 are entangled in the expression of secondary prominences in BP. Distinct from the prominence discussed so far, secondary prominences are realised in other positions than the stressed syllable. They signal the beginning of stress groups.

For applying the normalisation technique, the labelling of the phoneme-sized segments within each V-to-V interval is a necessary step that can be done manually or automatically (the *EasyAlign* tool developed by Goldman, 2011 delivers both phoneme-size boundaries and labels from an audio file. This tool was recently adapted to work on BP). As explained above, the normalised duration peaks can be used to define the right end of the stress groups in a right-headed language such as BP at this level. This allows both to count the number of phonetic syllables within the stress group as well as its duration automatically, which saves time and is useful for research on rhythm typology.

In fact, O'Dell and Nieminen (1999) showed that a tendency towards stress-timing (or syllable-timing) can be estimated from the ratio between the intersect and the slope of the linear regression line predicting stress group duration from the number of phonetic syllables within this group (see Barbosa et al., 2009 for an application to evaluate the rhythmic differences between European and Brazilian Portuguese). Stress-timing concerns the alleged sensation that phrasally stressed syllables occur regularly in time, whereas syllable timing concerns the alleged sensation that syllables occur regularly in time. The literature on rhythm typology is very large, but some reviews on the theme can be found to get started (e.g., Barbosa, 2000, 2006; Bertinetto, 1989).

Speech rate is also a variable that needs to be taken into account in rhythm research. It can be defined either as the number of phonetic or as the number of phonological syllables per second. Both speech rate increase and decrease affect the syllable-sized durations throughout the utterances as shown by Barbosa (2006, 2007) for BP. That's why speech rate needs either to be controlled (in that case it is a to-be-controlled variable) for not influencing the results or it needs to be manipulated (in that case it is an independent variable) to study its effects on the corpus under study. Meireles and Barbosa (2008) have evaluated the possible contribution of speech rate increase for explaining the emergence of penultimate from antepenultimate lexical stress patterns in BP.

Pausing is another important component of the rhythmic structure of an utterance. A pause can be realised with a silent interval (silent pause) or with a lengthened V-to-V unit not followed by a silence (filled pause). Pause is a sensation of break caused by these two acoustic possibilities, among others. Pause can also signal a hesitation, when it is called a hesitative pause. Merlo (2012) has recently demonstrated that hesitation and hesitative pauses help maintain fluency during narrative and descriptive instances. Pause can also be a signal of a difficult in production, as in the case of dysarthria. Besides have shown that the longitudinal study of pausing reveals the benefit of therapy in dysarthric speech, the work by Vieira (2007) also revealed another striking aspect of pausing in pathological speech. Even though the number of silent pauses in dysarthric speech is higher than the number of silent pauses in the control group, the hierarchy of these pauses, revealed by the statistical distinction among their duration, signals the underlying linguistic structure also highlighted by the control group.

Production and perception mechanisms of rhythmic structure were recently studied in an integrative way by Barbosa & Silva (2012). They demonstrated that the rate and height of V-to-V normalised duration peaks, associated to speech rate explain up to 71 % of the variance of listeners' judgments about differences in manner of speaking of three BP subjects.

To sum up, in this section the roles of the prosodic functions of prominence and boundary to rhythm research were presented. To help revealing them in the production domain, the phonetic syllable was defined. The V-to-V normalised duration values throughout the utterance define the rhythmic structure associated to this utterance. This structure is characterised by a sequence of duration peaks of differing degrees which contributes to the perception of different degrees of prominence, secondary prominences and boundary strength. Pausing is an integral part of this rhythmic structure that can also be revealed by the same procedure. The alleged regular succession of phonetic syllables and phrasally stressed phonetic syllables was implicated in the definition of syllable- and stress-timing in rhythm research. Differences in the rate and degree of boundary and prominence of these variables explain differences in perceived rhythm.

5. Conceptual and methodological aspects in intonation research

The word "Manuel" in the example given in Fig. 1 is perceived as prominent by the listeners even though there is no duration peak in the word. In fact, a rising F_0 contour within the word signals to the listeners the importance of this piece of information. The F_0 contour for the utterance can be seen in Fig. 3, where the rising contour is represented by the symbol LH.

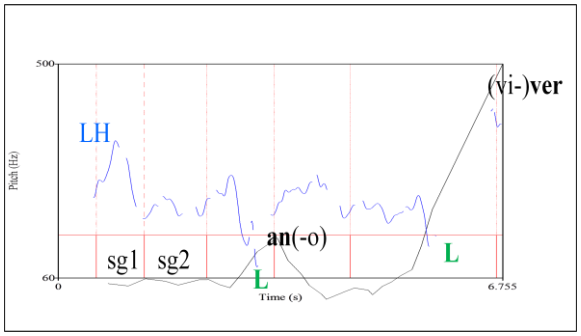

Figure 3: F_0 contour superposed to the V-to-V normalised duration contour of the same utterance of Fig. 1."sg1" and "sg2" signal the first two stress groups. The first one ends at the syllable "tra" in the word "entrado". The second one ends at the syllable "tei" in the word "mosteiro".

This clearly tells that prosody perception in BP depends on at least two acoustic parameters: syllable duration and F_0 movement. In fact, it depends on all acoustic parameters that signal prosodic information, including intensity, voice quality and even vowel quality (e.g., the lower F1 value of the last /a/ of "papa", pope, also signals the penultimate stress pattern). It is the work of the experimenter to determine which parameters contribute more to perceived stress.

F_0 patterns also signal the prosodic functions of prominence and boundary. At strong syntactic boundaries it is common that both F_0 and duration signal the corresponding prosodic boundary (Barbosa, 2008). This can be seen in Fig. 3, where the two main peaks of normalised duration, in "ano" and "viver", are accompanied by low levels of F_0, indicated with the L symbol.

Maybe because of the relevance of F_0 movements in signalling prominence and boundary in well-studied languages such as English, the term "intonation" is closely related to the term "prosody". That is why, before continuing it's necessary to say some words in this respect.

Hirst and Di Cristo (1998, p. 1-44) consider "prosody" as the general term including the lexical and post-lexical domains. For them, intonation is the study of the abstract relations in the post-lexical domain, independently of the acoustic parameter that signal these relations. In this sense intonation embraces the study of pitch accent and boundary tone patterning, as well as the study of duration patterns throughout the utterances.

Another possible approach stems from studies on

prosody perception and relies on the effects associated with the sensation of pitch, duration and loudness. For this approach, "prosody" is also the general term embracing the lexical and post-lexical domains, but "intonation", on the other hand, is restricted to the analysis of pitch variation throughout the utterances. Because the physical parameter that primarily controls the pitch sensation is F_0, the phonetic studies of intonation in this approach analyse the F_0 patterns throughout the utterances. It is this sense of intonation that we are using here. In this approach, "rhythm" is independent of "intonation" because it relies on the study of perceived syllable duration through the analysis of its main correlate, observed duration, as already depicted in the preceding section. Let's present some key concepts in intonation research.

Pitch accent is the intonation-related term for a prominence signalled by a F_0 movement, whereas the sensation of break is signalled by a boundary tone. Thus, pitch is not a synonym of F_0 peak or valley: it is a sensation that only can be evaluated by perception tests with real subjects. It cannot be measured in an objective way. Fig. 3 illustrates an F_0 movement perceived as a pitch accent in BP. The movement has a rising shape (LH) and is followed by two low boundary tones (L). These two low tones in BP fulfil the function of signalling terminality, as we will see later in this section. The rising movement is defined with relation to alignment of the rising part of the contour with the stressed syllable, as can be seen in Fig. 4, where the LH contour rising is entirely realised within the stressed syllable "lhões". Annotation of intonation-related prosodic functions is an important step to the study of intonation.

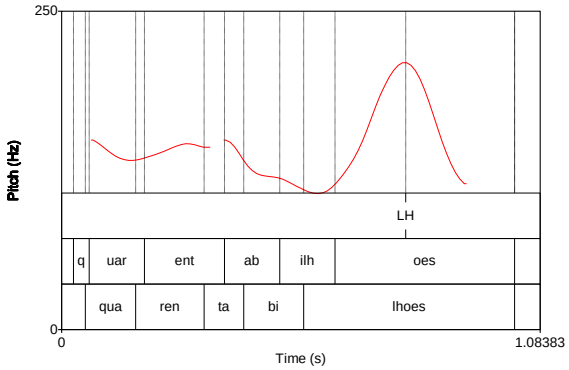

Figure 4: Illustration of the rising contour LH on the word "bilhões" from Lucente (2008).

Annotation systems such as ToBI, although largely adopted by researchers of American English (Silverman et al., 1992), German (Reyelt et al., 1996), and Spanish (Beckman et al., 2002), did not prove consistent across labellers (Wightman, 2002). By asking them to annotate pitch accent type by ear, the ToBI annotation procedure mixed up the roles of form and function in shaping intonation patterning (Hirst, 2005). To avoid this, the best solution is to only ask the listeners to indicate whether a word is prominent or not, and whether a word precedes a prosodic break or not. After this phase, labels are assigned by examining the movement of the F_0 with relation to the stressed vowel (or stressed syllable). This was done by Lucente (2008) for studying focus in BP with the proposal of the DaTO system of intonation annotation. Recently, she extended the analysis to examining the relation between pitch accents and information status (Lucente, 2012). Examples of contour labels from the DaTO system can be seen in Figs. 4 to 9.

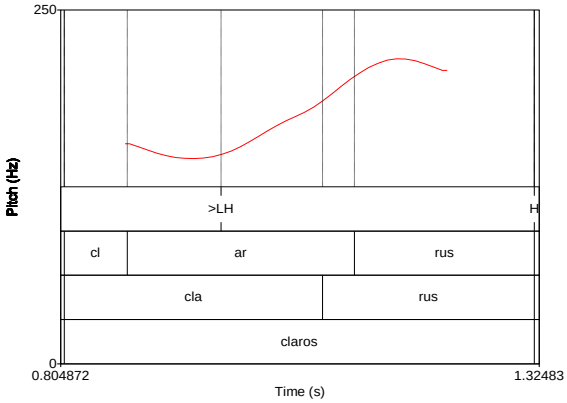

Figure 5: Illustration of the late rising contour >LH on the word "claros" from Lucente (2008).

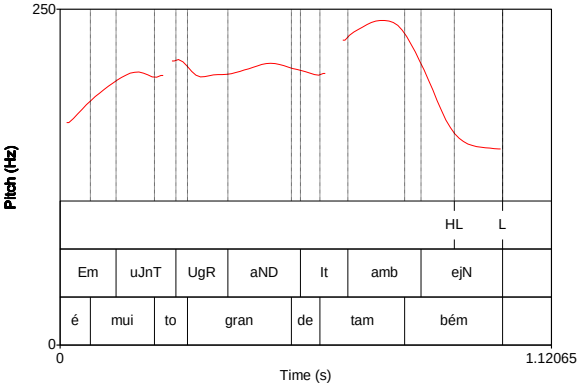

Figure 6: Illustration of the falling contour HL on the word "também" from Lucente (2008).

Figs. 4 and 5 illustrate the contrast between the rising and late rising contours. The F_0 peak occurs after the lexically stressed vowel in the latter case, whereas it occurs during the lexically stressed syllable in the former case. This contrast is similar to the one between the falling and late falling contours shown in Figs. 6 and 7. Observe in Fig. 6 that in the HL contour, the low level of F_0 is attained during the lexically stressed syllable by a sharp fall from a higher position. This sharp fall is delayed in the late falling contour exhibited in Fig. 7 where the lowest part of the F_0 contour levels out during the post-stressed syllable of the word "caras".

These differences are known as differences in tonal alignment. Recent work on intonation has shown that tonal alignment with respect to the syllable is a crucial component of the intonation system of a language (see Xu, 2005).

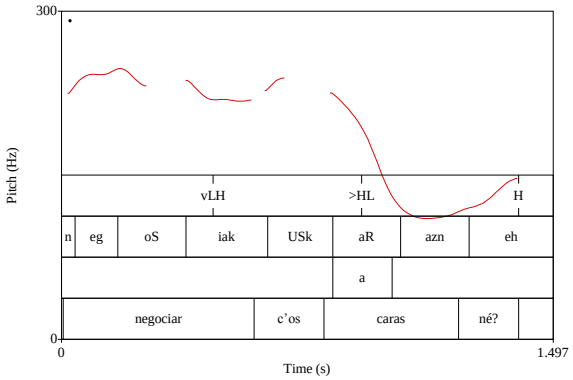

Figure 7: Illustration of the late falling contour >HL on the word "cara" from Lucente (2008).

The contours illustrated here are used by the speaker to signal prominence of the words onto which they are realised. Boundary tones are used to signal prosodic boundaries. Fig. 8 shows the realisation of a low boundary tone (L) in spontaneous speech, also illustrated in Fig. 3 in read speech. Low tones signal terminality in several Indo-European languages, although research about dialectal variability has shown that this picture is far from being simple (see Grabe, 2004 for prosodic variation in British English, where, in Newcastle English, almost 17 % of the declaratives are realised by a final high tone).

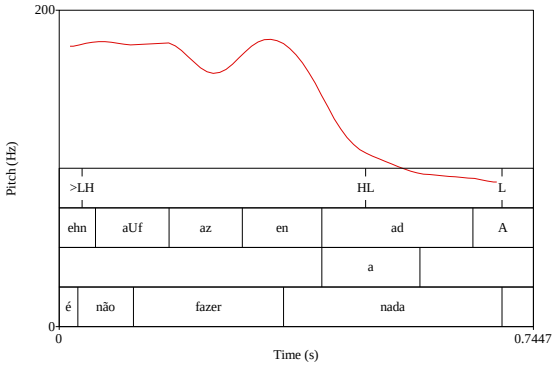

Figure 8: Illustration of the low tone contour L at the end of the word "nada" from Lucente (2008).

Fig. 9 signals a high contour tone (H) in standard German storytelling. This high tone at the end of the utterance signals the listeners that there is more to come: that is, high tones in German signal non-terminality. In the same speaking style, non-terminal boundaries signalling the continuation of a story are realised by a rising-falling contour in BP, as shown in Fig. 10 in two positions during the narration. Furthermore, Barbosa et al. (2011) showed that, in contrast with Standard German, during storytelling, BP –speaking subjects often maintain a high F_0 level between prominent words, as can be seen in Fig. 10 within the dotted ellipsis. In this excerpt, the speaker repeats part of the information she just gave, that the monk did not accustom with the routinely activities of the monastery. The stressed syllable of the word "acostumava" is extremely lengthened.

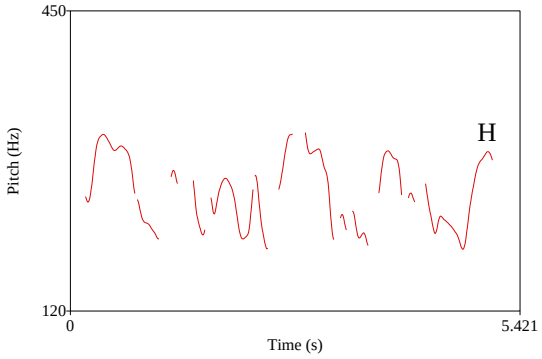

Figure 9: Illustration of the high tone H contour at the end of the word "gewöhnt" in a female speaker of read Standard German from Barbosa et al. (2011).

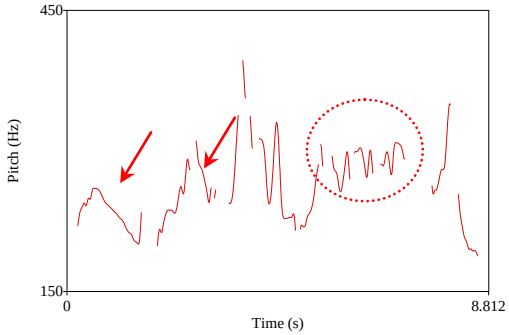

Figure 10: Continuative contours in the words "ele" and "que" (first two arrows from left to right) and high F_0 register (dotted elipsis) in the passage "ele não se acostumava com a rotina do [...]" during the narration of a BP female speaker

There is a close similarity between F_0 shapes for signaling yes-no questions and continuation of dialogue turns in BP. Both are signalled by rising falling contours whose difference relies on the alignment of the rising part of the contour. Fig. 11 shows the rising-falling shapes in the same word "seguida" from the expression "em seguida" (in the following) realised by a male speaker from the State of São Paulo. It can be seen that the continuative contour rightwards is relatively low during the lexically stressed syllable with almost the entire rising realised during the post-stressed syllable /da/. On the other hand, the rising of the yes-no question contour leftwards resides in the stressed syllable /gi/. The difference in degree between the F_0 peaks in the two contrasting contours is related to the degree of emphasis the speaker put in the continuative contour. He could have realised the yes-no question with more emphasis, if necessary for communicative reasons. The crucial acoustic component for distinguishing yes-no questions from continuative turns in BP is the delay of the rising part of the contour in the second case.

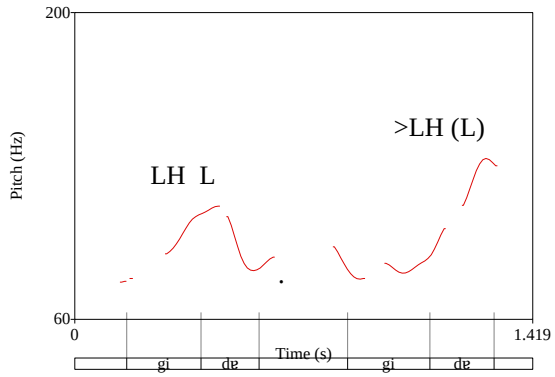

Figure 11: Contrast between yes-no question LH L (left) and continuative >LH L (right) contours in a male speaker of BP in the the word “seguida”.

To sum up, intonation research needs an annotation system related to the classical prosodic functions of prominence, boundary marking and discourse event marking to be appropriately carried out. Recent research strongly suggests that annotation should relate F_0 contours to landmarks in the syllable. Defined functionally, pitch accents and intonational breaks can be adequately studied. Tonal alignment is a crucial element for distinguishing the contour types. Terminality and non-terminality are signalled by boundary tones which are different cross-linguistically. Intonational differences across languages can also be related to the way the F_0 curve between prominent intonational events is realised.

6. Expressive speech research

The relevance of the vocal expression to signal affect was recognised at least as early as the XIXth century (Darwin, 1872 apud Scherer, 1986, p. 143). Scherer (1981) showed that naïve judges are more precise in assessing vocal than facial expression. The problem is to find out acoustic correlates for explaining this successful perceptual recognition. F_0 is certainly one of these parameters, at least as far as the study of high-arousal emotions are concerned (Scherer, 1986, p. 144; Frick, 1985, p. 418).

Emotion is only one of the possible affective states carried by the speech signal. Affect also includes mood, attitudes and interpersonal stances, preferences, and affective dispositions, as proposed by Scherer (1984). In comparison with the other affects, emotion is short in duration, it is more intense in terms of body responses, it triggers a simultaneous behaviour in other parts of the organism, and it is synchronous with the event that triggered the emotional behaviour. In everyday life, all affects are usually present in a single utterance. That is why the area of research dealing with affect in speech is called expressive speech research.

Several acoustic-prosodic parameters can be extracted from an utterance, which are relevant for expressive speech studies. The most used are F_0, long-term average spectrum (LTAS), syllable-size duration and speech rate, as well as voice quality. Statistical descriptors such as mean, standard-deviation and skewness are used to evaluate the differences across different affects, such as the work on attitudes carried out by Moraes and colleagues in BP (Moraes, 2011; Moraes et al., 2010; Rilliard et al., 2012).

Another research approach in expressive speech studies is the evaluation of changes in expressiveness during sequences of utterances during conversations, as was done by Barbosa (2009) for BP. In this study, where circa 200 utterances extracted from a radio show were examined, an experiment designed to study the relation between perceived and produced expression was run out. The evaluation of the utterances was done by a set of judges and the prediction of the evaluation rates from a set of acoustic parameters. For predicting the rates, the set of utterances was split into two subsets, the training subset with 130 randomly chosen utterances from 12 subjects, and the test subset with the remaining 76 utterances. The training subset was evaluated by 12 judges, all of them undergraduate students of the first year in Linguistics. Four affect dimensions were evaluated by all judges in different days in two weeks. The dimensions were activation, involvement, valence, and dominance. The use of a dimensional approach in expressive speech research avoids the inter-subject variation in judgment if affective words are used due to idiosyncratic experience with each affect. Dimensional analysis has its limits: as it has been used, it could mask the dynamical aspects of affect change or rely only on the dimensions analysed to understand affect evaluation (see Scherer, 2000 for a criticism). These two drawbacks were avoided by using a Principal Component Analysis (PCA) to discover the main axes of variation in judgment when combining the dimensions chosen for analysis. All dimensions are evaluated within a 7-point differential semantic scale between two poles. Activation is a value between relaxed/calm and agitated/stimulated. Valence is a value between pleasant and unpleasant. Involvement is a value between involved and non-involved, and dominance is a value between under-control and submissive. This latter dimension was not reliably evaluated across judges and it was discarded.

Two factors in the PCA explained 97 % of the variance of the judgments, where factor 1 was related to arousal and explained 90 % of the judgments. To infer the judges’ evaluation median rates for all utterances and dimensions, five classes of acoustic parameters were extracted: F_0, F_0first derivative (dF_0), intensity, spectral tilt (SpTt), and Long-Term Average Spectrum (LTAS). Up to four statistical descriptors were used for each class, producing twelve acoustic parameters: F_0 median, inter-quartile semi-amplitude, skewness, and 0.995 quantile; dF_0 mean, standard-deviation, and skewness; intensity skewness; spectral tilt mean, standard-deviation, and skewness; and LTAS standard-deviation. Spectral tilt is a correlate of vocal effort and was set to the difference of intensity in dB between the bands 0−1250 Hz and 1250−4000 Hz.

The spectral tilt descriptors and the dF_0 mean predict the new arousal dimension (factor 1) of the judgments’ evaluations, with a correlation of 67 %. If these

predicted-from-acoustics values are arranged chronologically in terms of the radio show participant, it is possible to detect changes in behaviour, as can be seen in Fig. 12.

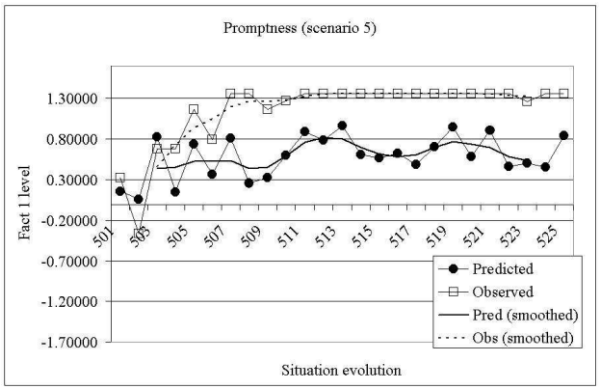

Figure 12: Predicted and observed values of arousal (promptness) in a scenario where the participant of a radio show is very irritated.

From utterances 501 to 505 the participant talks to his daughter and from utterance 506 on, to the radio presenter. The observed contour shows, because evaluated by the judges, a saturation to a maximum level of arousal. This is not the case of the predicted-from-acoustics contour, which shows a trend to higher levels of arousal with some oscillations. The predicted levels are entirely based on acoustic parameters, contrary to the observed rates. These latter are also dependent on other influences, such as the semantic weight of the lexical items. In this situation, it is likely that the judges inferred the reasons for the participant's rage and decided to choose maximum levels of arousal, given the lexical items used by the participant. Nevertheless, the predicted values can be used to detect subtle changes in expression, such as the increasing of arousal from utterances 508 to 511.

The application to automatic detection of expressiveness is immediate.

7. Acknowledgements

The works referred to here were supported by grants from CNPq (301387/2011-7, 300371/2008-0, 400280/2009-4 and 490726/2008-9).

8. References

Arantes, P. (2010). Integrando produção e percepção de proeminências Secundárias numa abordagem dinâmica do ritmo da fala. Doctoral Thesis. State University of Campinas.

Barbosa, P. A. (submitted). Conhecendo melhor a prosódia: aspectos teóricos e metodológicos daquilo que molda nossa enunciação.

Barbosa, P. A. (2009). Detecting changes in speech expressiveness in participants of a radio program. In *Proc. of Interspeech 2009 - Speech and Intelligence*, Brighton, UK. London: Causal Productions, pp. 2155 --2158.

Barbosa, P. A. (2008). Prominence- and Boundary-Related Acoustic Correlations in Brazilian Portuguese Read and Spontaneous Speech. In *Proc. Speech Prosody 2008*, Campinas, pp. 257--260.

Barbosa, P. A. (2007). From Syntax to Acoustic Duration: a Dynamical Model of Speech Rhythm Production. *Speech Communication*. 49, 725--742.

Barbosa, P. A. (2006). *Incursões em torno do ritmo da Fala*. Campinas, Brazil: FAPESP/Pontes Editores.

Barbosa, P.A. (2000). "Syllable-timing in Brazilian Portuguese": uma crítica a Roy Major. *D.E.L.T.A*, 16 (2), 369--402.

Barbosa, P. A. (1996). At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration: emphasis on segmental duration generation. *Cadernos de Estudos Lingüísticos*. 31, pp. 33—53.

Barbosa, P. A. (1994). Caractérisation et géenération automatique de la structuration rythmique du français. Thèse de Doctorat de l'INPG. Institut National Polytechnique de Grenoble, France.

Barbosa, P.A.; Mixdorff, H.; Madureira, S. (2011). Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese. In *Proc. of Interspeech 2011*, Florence, Italy, pp. 2065-2068.

Barbosa, P. A.; Viana, M. C.; Trancoso, I. (2009). Cross-variety Rhythm Typology in Portuguese. In *Proc. of Interspeech 2009 - Speech and Intelligence*. Brighton, UK. London: Causal Productions, pp. 1011 --1014.

Barbosa, P. A.; Silva, W. (2012). A New Methodology for Comparing Speech Rhythm Structure between Utterances: Beyond Typological Approaches. In H. Caseli et al. (Eds.), *PROPOR 2012*, *LNAI 7243*, Springer: Heidelberg, pp. 329--337.

Beckman, M. et al. (2002). Intonation across Spanish, in the Tones and Break Indices framework. *Probus*, 14, PP. 9--36.

Bertinetto, P. M. (1989). Reflections on the dichotomy "stress"- vs "syllable-timing". *Revue de Phonétique Appliquée*, 91–92–93, pp. 99–-130.

Beveridge, W.I.B. (1957). *The Art of Scientific Investigation.* New Jersey, USA: The Blackburn Press.

Bolinger, D. (1986). *Intonation and Its Parts: Melody in Spoken English*. Stanford: Stanford University Press.

Botinis, A.; Granström, B.; Möbius, B. (2001). Developments and paradigms in intonation research. *Speech Communication*, 33, pp. 263-296.

Bunge, M. (1998). *Philosophy of Science. From Explanation to Justification.* New Brunswick, NJ: Transaction Publishers.

Byrd, D.; Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*. 26, pp. 173--199.

Classe, A. (1939). *The Rhythm of English Prose*. Oxford: Blackwell.

Crawley, M. J. (2005). *Statistics. An Introduction using R.* Sussex, UK: John Wiley & Sons, Ltd.

Edwards, J.; Beckman, M. E; Fletcher, J. (1991). The articulatory kinematics of final lengthening. *J. Acoust.*

Soc. Am. 89 (1), pp. 369--382.

Fleck, L. (1992 [1935]). Observation scientifique et perception en général. In J.-F. Braunstein (Org.), *L'Histoire des sciences*. Paris: Librairie Philosophique J. Vrin, pp. 245--272.

Fónagy, I. (1986). Les langages de l'émotion, *Quaterni di semantica*. 7/2, pp. 305--318.

Frick, R. W. (1985). Communicating emotion: the role of prosodic features. *Psychological Bulletin.* 97 (3), 412—429.

Goldman, J.-P. (2011) *EasyAlign: an automatic phonetic alignment tool under Praat*. In *Proc. of the Interspeech 2001*. Florence, Italy, pp. 3233--3236.

Grabe, E. (2004). Intonational variation in urban dialects of English spoken in the British Isles. In P. Gilles & J. Peters (Eds.), *Regional Variation in Intonation*. Tübingen: Niemeyer, pp. 9--31.

Hirst, D. (2005). Form and Function in the Representation of Speech Prosody. *Speech Communication*. 46, pp. 334—347.

Hirst, D.; Di Cristo, A. (1998). *Intonation systems: a survey of twenty languages*. Cambridge: Cambridge University Press.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Massachussets: MIT Press.

Lucente, L. (2012). Aspectos Dinâmico-Funcionais do Foco e da Curva Entoacional do Português Brasileiro. Doctoral Thesis. State University of Campinas.

Lucente, L. (2008). Dato: Um Sistema de Notação Entoacional do Português Brasileiro baseado em Princípios Dinâmicos. Ênfase no Foco e na Fala Espontânea. Master's Thesis. State University of Campinas.

Massini, G. (1991). A Duração no estudo do acento e do ritmo em português. Master's Thesis. State University of Campinas.

Meireles, A.; Barbosa, P. A. (2008). Lexical reorganization in Brazilian Portuguese: An articulatory study. *Speech Communication*. 50, pp. 916--924.

Merlo, S. (2012). Dinâmica temporal de pausas fluentes e hesitações na fala semi-espontânea. Doctoral Thesis. State University of Campinas.

Moraes, J. A. (2011). From a prosodic point of view: remarks on attitudinal meaning. In H. Mello et al. (Eds.), *Pragmatics and Prosody. Illocution, modality, attitude, information patterning and speech annotation.* Firenze: Firenze University Press, pp. 19--37.

Moraes, J. A. (1998). Intonation in Brazilian Portuguese. In D. Hirst, & A. Di Cristo (Eds.), *Intonational Systems: a Survey of Twenty Languages*. Cambridge: MIT Press, pp. 179--194.

Moraes et al., (2010). Multimodal perception and production of attitudinal meaning in Brazilian Portuguese In *Proc. of the Speech Prosody 2010 Conf., pp.*

Reyelt, M et al. (1996). Prosodische Etikettierung des Deutschen mit ToBI. In D. Gibbon (Ed.), *Natural Language and Speech Technology, Results of the third KONVENS conference*, Berlin: Mouton de Gruyter, pp. 144—155.

Rilliard, A. et al. (2012). Prosodic analysis of Brazilian Portuguese attitudes. In *Proc. of the Speech Prosody Conf.* Shanghai, pp.

Sauvanet, P. (2000). *Le Rythme et la raison. Rythmologiques.* Paris: Éditions Kimé.

Scherer, K. R. (2000). Psychological models of emotion. In J. C. Borod (Ed.), *The Neuropsychology of emotion*. New York: Oxford University Press, pp. 137--162.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin.* 99 (2), 143--165.

Scherer, K. R. (1984). On the nature and function of emotion: a component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion*. Hillsdale, NJ: Lawrence Erlbaum, pp. 293--318.

Scherer, K. R. (1981). Speech and emotional states. In J. Darby (Ed.), *Speech evaluation in psychiatry*. New York: Grune & Stratton, pp. 189--220.

Scott D.R. (1980). Duration as a clue to the perception of a phrase boundary. *Journal of the Acoustical Society of America*. 71, 996 –1007.

Scott, S. K.; Wise, R. J. S. (2003). PET and fMRI studies of the neural basis of speech perception. *Speech Communication*, 41, pp. 23–34.

Silverman, K. et al. (1992). ToBI: a Standard for Labeling English Prosody. In *Proc. of the 2nd International Conference on Spoken Language Processing*. Banff, Canada, 2, pp. 867--870.

Tabain, M. (2003). Effects of prosodic boundary on /aC/ sequences: articulatory results. *J. Acoust. Soc. Am.* 113 (5), pp. 2834--2849.

Traunmüller, H.; Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am.* 107, 3438--3451.

Vieira, J. M. (2007). Para um estudo da estruturação rítmica na fala disártrica. Doctoral Thesis. State University of Campinas.

Whitehead, A. N. (1919). *An Enquiry concerning the Principles of Natural Knowledge.* Cambridge, UK: Cambridge University Press.

Wightman, C. (2002). ToBI or Not ToBI? In *Proc. 1st Internat.Conf. on Speech Prosody*. Aix en Provence.

Wightman, C.W. et al. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *J. Acoust. Soc. Am*., 91, pp. 1707--1717.

Wunenburger,J.-J. (1992). *Les rythmes. Lectures et theories.* Paris: L'Harmattan.

Xu, Y. (2011). Speech Prosody: A Methodological Review. *Journal of Speech Sciences*, 1(1), pp. 85--115.

Xu. Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*. 46, pp. 220--251.