Speech Prosody in Musical Notation: Spanish, Portuguese and English

Antônio R.M. Simões[1], *Alexsandro R. Meireles*[2]

[1] University of Kansas, Lawrence, USA
[2] Federal University of Espírito Santo, Vitória, Brazil
asimoes@ku.edu, meirelesalex@gmail.com

Abstract

This study uses musical notation to describe speech prosody in connected speech in Brazilian Portuguese and Mexican Spanish, using English as a comparison where needed. Through this research we establish the basis on which to expand our future work on speech prosody, from methodology to data collection and analyses, and then make initial observations regarding potentially significant prosodic patterns. This study shows that musical notation can inform us about: 1) the pitch ranges of the speakers in connected speech; 2) speech rate; 3) patterns of moraic and non-moraic syllables; 4) syllable timing; 5) intonation patterns, especially speakers' tessitura. The methodology that we have developed in this exploratory study may help solve unpredictable patterns of speech prosody, especially in regards to intonation, and consequently lead to the improvement of current speech prosody models.

Index Terms: prosody, speech rhythm, timing, intonation, Spanish, Portuguese, musical notation

1. Introduction

One of the most resourceful aspects of this study is the use of musical theory in the transcription of speech for empirical analyses, as illustrated in Figure 1. As the authors continue to progress in the study of the rhythmic patterns in Spanish and Portuguese, the promising potential of this research has become more evident. The results of this research will continue to develop our understanding of how speech prosody works in natural languages. Furthermore, the methodology used here may help clarify or solve unpredictable patterns of speech prosody, especially with respect to speech timing and intonation, and consequently lead to the improvement of current speech prosody models.

Although musical theory has been used in the past and forms the basis of the British School tradition in Linguistic Studies of language intonation [1] [19] [27], the focus on an empirical study of timing in discourse through music theory has not been fully explored. These earlier studies of intonation in Britain were done for pedagogical purposes. Later, in the 1960s, their pedagogical tools gained the interest to other British scholars [11] [13] [18] [23] [26] who gave them an empirical basis for the studies of intonation, but without abandoning its didactic side. In this study, we are interested in the didactic side of prosodic studies, but the main interest is on the use of empirical data and how the results affect the existing theoretical frameworks and speech prosody models.

Brazilian Portuguese, for example, contrary to what has been observed in Spanish, has shown limited predictability of intonation patterns according to João Moraes [24] [25] [and personal communication]. Preliminary analyses of rhythmic patterns that the authors have so far carried out have given insights that could lead to the improvement of predictability in intonation patterns. Furthermore, current models for intonation

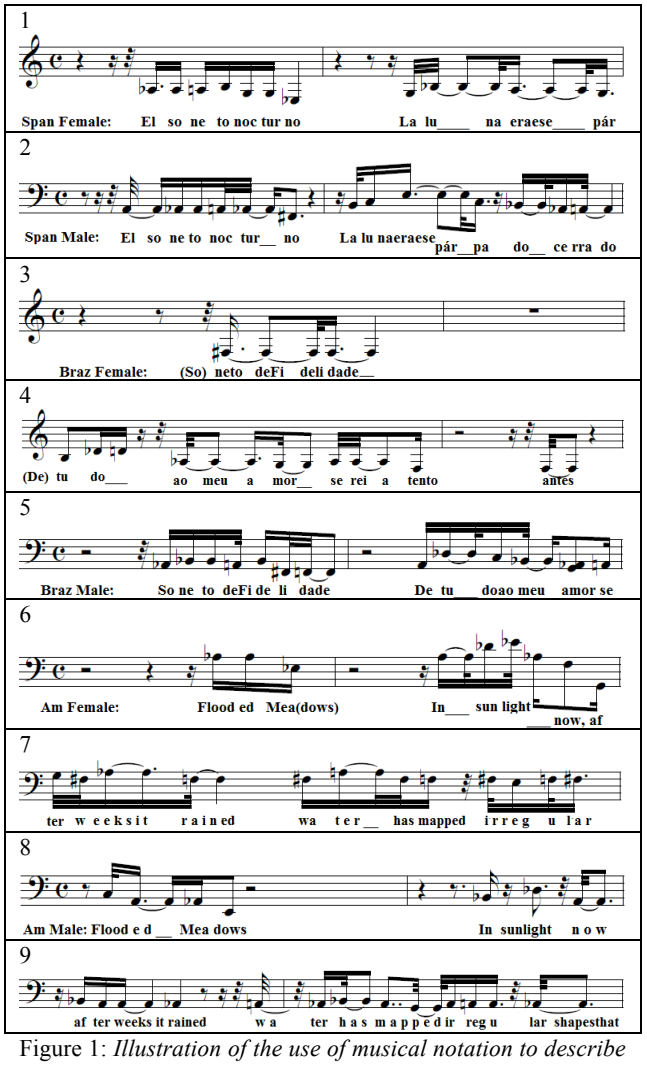

Figure 1: *Illustration of the use of musical notation to describe the timing and pitch features of sonnet readings by a Spanish female (1), Spanish male (2), Brazilian female (3, 4), Brazilian male (5), American female (6, 7), and American male (8, 9). Note that the words in parentheses indicate errors in the transcription due to creaky voices.*

are still limited in their performance as discussed in [3] [4] [5] and [6]. Research done using musical notations of large-scale spontaneous recordings can reveal rhythmic trends which,

 doi: 10.21437/SpeechProsody.2016-44

combined with intonation, may improve our understanding of how intonation works and contribute to the development of speech intonation models such as [4] [5] [6] [34] [35] [37] [39].

Likewise for intonation, studies of speech rhythm or speech timing in natural languages were done initially for pedagogical purposes, such as Kenneth Pike's seminal study [32]. Pike's study is still helpful in the teaching of languages like Spanish and English, although the notions of syllable-timed and stress-timed rhythms may be difficult to prove empirically. For a survey of recent studies on speech rhythm see [28] and [29]. We hope that musical notation will bring insights to better predict when to expect certain rhythmic and intonation patterns in spontaneous speech.

Since the early 20th century, the development of prosodic studies reflects the elusive nature of prosody. The disparity of approaches to analyze prosody and the almost chaotic proliferation of an endless terminology in prosodic studies to date confirms the difficulty one finds when studying any area of prosody, especially speech rhythm and intonation. Studies carried out within the main schools of prosodic studies, such as the British School, the Dutch School, the structuralist tradition in the US, the generativist tradition in the US as well as the generativist trend in Aix-en-Provence, France, have developed different views of how speech prosody works. To minimize these differences, there have been attempts to standardize the transcription of intonation. For instance, the Metric-Autosegmental tool for labeling intonation, now called MAE-ToBI, was promoted by workshops in 1991, 1992, 1993, and 1994 [8] [9] to resolve these differences. But then, scholars of languages other than English created new versions of ToBi for other languages, including Spanish and Portuguese. They were important steps towards developing a uniform view in intonation studies, but even those efforts did not halt the disagreements, multiplication of views, and the multiplication of terminology.

This investigation aims at staying as close as possible to the notation system in the musical tradition, although it takes into account the richness of the work accumulated so far through other systems of speech notation.

2. Methodology

This study uses speech recordings of native speakers of Brazilian Portuguese, Mexican Spanish, and American English. English data are used English as interface. We recorded six subjects, three pairs of one male and one female, in a sound proof booth, at the EGARC Language Laboratory at the University of Kansas. The speakers read five times one sonnet in their native languages. Of the five readings, one reading was chosen for this study, usually the third or fourth reading, depending on the subject's reading performance. The sonnets were written by Thom Gunn ("Flooded Meadows"), Felipe Benítez Reyes ("El soneto nocturno"), and Vinícius de Moraes ("Soneto de Fidelidade"). Each subject became well-acquainted with the content of the sonnet before reading it five times.

After the recorded passages were selected, the authors used the softwares Ableton Live and Finale for automatic musical notation. This automatic notation is the first step in the process of using musical notation. The final transcription is done through visual, auditive, and manual verification of the recordings to further improve the final transcriptions. This manual verification also provides information that is very useful to improve automatic detection of speech prosody, especially timing.

Table 2: *The FREQ Procedure applied to the reading of a sonnet by a female speaker of English. The small letter "d" stands for dot or dotted. For instance, regular notes and rests have no ds; 4.1d means a one-quarter note or a one-quarter rest. The rows for notes and rests mean frequency, percent, row percent and column percent.*

Table 1 of type by total

Controlling for language=English, gender=female

	Total											
	1.0, 2.0, 4.0, 6.0, 8.0, 16.0, 32.0 = regular full note or full rest; 4.1d, 8.1d, 8.2d, 16.1 means one-dotted or two-dotted x-note or x-rest.											
type	**1.0**	**2.0**	**4.0**	**4.1**	**6.0**	**8.0**	**8.1**	**8.2**	**16.0**	**16.1**	**32.0**	**Total**
note	0 0.00 0	0 0.00 0	5 2.98 3.91 50.0	0 0.00 0 .	0 0.00 0.00 .	14 8.33 10.94 73.68	8 4.76 6.25 100.0	0 0.00 0 .	40 23.81 31.25 80.0	24 14.29 18.75 100.0	37 22.02 28.91 68.52	128 76.19
rest	0 0.00 0	3 1.79 7.50 100.0	5 2.98 12.5 50.0	0 0.00 0.00 .	0 0.00 0.00 .	5 2.98 12.5 26.32	0 0.00 0.00 0	0 0.00 0.00 .	10 5.95 25.00 20.0	0 0.00 0.00 0	17 10.12 42.50 31.48	40 23.81
Total	0 0.00	3 1.79	10 5.95	0 0.00	0 0.00	19 11.31	8 4.76	0 0.00	50 29.76	24 14.29	54 32.14	168 100.0

Table 2: *The estimated errors in the musical notation, for the American, Mexican, and Brazilian speakers*.

	Total Events	Frequency of Errors	Percent of Errors/Events
Am Male	252	0	0%
Am Female	168	17	10%
Span Male	205	0	0%
Span Female	208	7	3.37%
Braz Male	210	0	0%
Braz Female	239	19	7.95%

Once the notation is concluded, the speech transcription is quantified by giving numerical values to notes and rests, which are the denominators in the musical notation, e.g. 4 was the value for a quarter note or rest. This quantification is essential for the organization of our data and the use of statistical tests in the search for speech patterns of speech timing, i.e. rhythm.

Data was entered in Excel and treated statistically using the SAS 9.4 package. Table 1 shows how the data was organized for initial analysis. This organization of the data helps to understand our corpus and establish research questions. Our work with the musical notation of speech prosody is still exploratory. And as such our research questions come up after the musical notation is done and its data organized according to the frequency of the speech events, as illustrated in Table 1. A second table, Table 2,

contains a summary of the number of speech events recorded and an estimate of the errors in the musical transcription of our six speakers.

3. Discussion

Musical notation, contrary to other transcription systems, is already universal by tradition. It can point out language behavior in a way that allows for rapidly transcribing and searching speech patterns. Music software packages are already designed to do most of the work for the automatic transcription of speech in musical notation. Furthermore and perhaps more importantly, musical notation is designed to represent both the dynamic and static events of music and by extension, speech prosody. This is so because we interpret the notes as static events, given the existence of micro-movements in speech. However, the duration of the notes is dynamic, since it represents time with exactitude. Even the dynamic intensity can be inferred in MIDI data through MIDI velocity, namely the greater the intensity, the greater the MIDI velocity.

Current automatic transcriptions are not perfect, but they can be improved by auditory and visual inspection of the automatic outputs for correction, once these automatic transcriptions are finished. By systematically correcting these outputs we may also improve the automation of musical notation, in addition to other contributions of this type of descriptive work. In this study this automation has been performing anywhere between 0% error to approximately 10% error, depending on the speaker gender and other sources of errors, as presented in Table 2.

Although our preliminary examination of our data has shown a number of potentially significant patterns, it is still premature to discuss all of them here. We will focus on the preliminary information gathered so far, namely the pitch range of the speakers in connected speech, speech rate, patterns of moraic and non-moraic syllables, syllable timing, and intonation patterns, especially the tessitura of our six speakers depicted in Tables 3 and Figure 2.

Table 3: *The tessitura of the six speakers. The darker the cell is, the higher the frequency (occurrence) of the notes. The numbers indicate precisely how many times the notes appear in the musical notation. The lower notes A, B and C did not show in the transcription, and were not included on this table.*

	Low										High
	G	D	E	F	G	A	B	C	D	E	F
Spn Fem		2	8	33	35	**52**	11	4	1	2	2
Spn Male				10	8	**83**	54	7	7	5	
Braz Fem			14	**72**	43	28	9	1	2		
Braz Male			1	31	28	**79**	25	4	5		
Eng Fem	1	5	22	**60**	8	27	3		1	1	
Eng Male			4	47	30	**80**	28	7	3		

The pitch ranges of the six speakers as shown in Table 3 and Figure 2 indicate not only the reach of each speaker's voices, but also their most common notes, which may indicate their most comfortable zone of speaking. These patterns are useful to determining a person's preferred level of speaking in terms of pitch range, and as a result they can indicate some difference in the speaking attitudes of speakers.

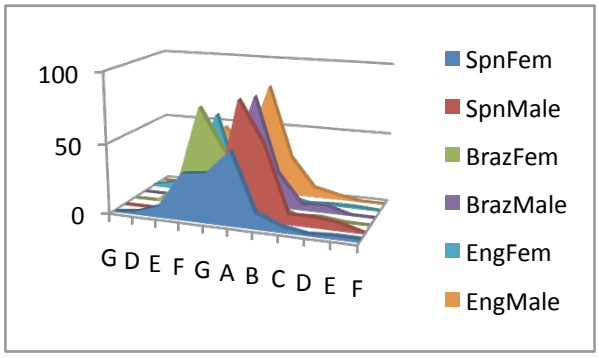

Figure 2: *The tessitura of the six speakers seen from a different view.*

Speech rate can quickly be determined by adding together notes and rests. The number of subjects is still limited, but the six ones that we analyzed cannot be differentiated by factors such as gender, as one can infer from Table 4.

Table 4: *Frequency of notes and rests of the six speakers. The letter "d" means "dotted note or rest," "N" means "notes," "R" means rests, "ST" means "sub-total," and "T" means "total." Three columns with one occurrence were deleted, but counted in the totals.*

		2	4	8	8	8	16	16	32	
		0 d	0 d	0 d	1 d	2 d	0 d	1 d	0 d	T
SpF	N	0	3	8	5	0	55	37	50	158
	R	2	5	8	0	0	16	0	19	50
ST		2	8	16	5	0	71	37	69	208
SpM	N	0	5	14	5	1	63	31	57	177
	R	0	4	7	0	0	7	0	10	28
ST		0	9	21	5	1	70	31	67	205
BrF	N	0	13	32	5	1	37	30	48	166
	R	4	16	11	0	0	17	0	20	70
	ST	4	29	43	5	1	54	30	68	236
BrM	N	0	4	10	4	1	65	25	68	177
	R	2	2	7	0	0	12	0	10	33
ST		2	6	17	4	1	77	25	78	210
EgF	N	0	5	14	8	0	40	24	37	128
	R	3	5	5	0	0	10	0	17	40
ST		3	10	19	8	0	50	24	54	168
EgM	N	0	11	28	10	4	48	31	61	194
	R	1	12	9	1	0	17	0	17	57
ST		1	23	37	11	4	65	31	78	251
T		12	85	153	38	7	387	178	414	1278

Several features may be indicative of bimoraic syllables like ligatures, dotted notes, dotted rests, as well as a larger pitch inflexion as observed in the word "light" that the American female speaker pronounced with the inflexion of the

high notes D-F-E (Figure 1.6). It is still early to determine how morae are rendered in the musical notation of speech, but these features may be the indicators. Sometimes, however, these same features indicate word stressed syllables.

In terms of syllable timing, our observations indicate that the six speakers alternate what can be considered syllable-timed rhythm with stress-timed rhythm. The Spanish speakers show a greater trend to syllable-timed rhythm, although the English speakers as well as the Brazilian speakers can also show some passages in syllable-timed rhythm. It was not possible to determine a clear relation of these different timings with specific syntactic structures, although we could for example observe that some seemingly staccato passages are clearly limited to certain readings, like the sonnet titles, as Figure 1.3 shows. The Spanish male speaker shows longer passages of staccato readings as Figure 3 below can attest. Curiously, the Spanish female shows a lot of pitch inflexions, which we think are typical of English. It may be due to her being a Spanish-English bilingual with strong although not dominant presence of English in her bilingualism. The Spanish male speaker is also bilingual with Spanish being his predominant language.

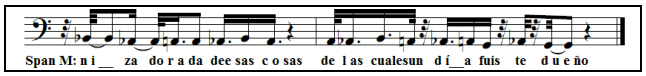

Figure 3: *An example of a staccato passage in the reading of a Mexican Spanish male.*

In Table 2, we consider the notes and rests as speech events. In order to estimate the errors, we divided the number of errors by the number of events, which gave us the percentage estimate. Whenever we saw an error in the automatic notation, we put it in parentheses, to indicate passages missed by the software. Then we counted the syllables missed, and in general each syllable is one error. Obviously this can change, especially in the case of moraic syllables.

As depicted in Table 2, musical notation errors occurred in the speech of the three female speakers. This should be expected because female voice is commonly characterized by creaky voices, which results in a technical flaw during software capture of these events, due to the usual lack of regular fundamental frequency in creaky voice.

4. Conclusions

We have argued in this study that musical notation is a system of transcription that can help solve many of the current problems that we encounter with transcription systems for speech prosody created in the last decades. One of the many advantages of musical notation is that it applies equally to all languages, contrary to what we have seen with the great efforts to create systems of prosodic transcriptions. Students in Linguistics and those interested in speech prosody can enroll in regular music classes to learn the sophistication and efficiency of musical notation.

This study is a first step for us, the authors, to use musical notation to gain insights into how speech prosody works. It was necessary to go through this initial exploratory step to establish a long term goal for the of use musical notation in our research in speech prosody. Having accomplished our initial goal, our next step is to increase the number of speakers as well as the number of languages to analyze.

Musical notation is an excellent means to understand and compare speech prosody across languages. The use of musical transcription may at first intimidate researchers who are not musicians. However, current music software systems can eliminate such intimidation. We relate more easily to musical notation of speech than we do to music, because we are more used to speech, especially in our native language.

After the first attempts to use musical notation, everything becomes easier to understand, even at some sophisticated levels of musical theory. For example, less frequent features may emerge in the process, but given the preciseness of what musical notation captures in music and speech, such less frequent features will make a lot of sense and bring additional insights to our understanding of language and speech. Furthermore, speech features obtained through musical notation are as quantifiable as everything else in musical notation. One of these less frequent features in this study was the identification of a few cases of appoggiaturas in the speech prosody of our speakers. After analyzing those few cases, we realized that they are in fact cases of appoggiaturas in speech and that some of the appoggiaturas point out to a possible relation of appoggiaturas to diphthongs and other speech events, as in the realization of the stressed vowel <o> in the word <a.mor> (*love*), in the recording of the Brazilian male speaker.

We are still testing these new grounds. There is still work to be done to find out the best ways to treat these data statistically. There may be, for example, associations between notes and rests in the musical notation in this study. We have done visual inspection of the patterns in rests and notes and noticed that there is a tendency for adjacent notes and rests to have the same values. Either the following rest repeats the value of the note or vice-versa.

We are still studying the possible statistical strategies that will help us to discover the significance of these associations. Our research questions continue to appear as we progress.

The next stage of our work will include experiments based on this current study and additional languages. We also need to conduct experiments to verify if it is possible to conform the diatonic scale to patterns of speech segments. In addition to such experiments, we will convert musical notes into equivalent pitch traces, i.e. fundamental frequency, in order to develop better notation system and improve the readership of our work, which is now limited to those interested in music.

5. Acknowledgements

The authors would like to thank Lesa Hoffman, at the University of Kansas (KU), for her kind assistance with our work on the statistical analyses. For our support in musical notation, we thank the musicians Vlad Geana and Michael Paul. To our colleague at KU, Jonathan Mayhew, our thanks for calling to our attention the poems by Thom Gunn and Felipe Benítez Reyes. We are very appreciative of the six speakers who kindly read the sonnets for this study, and to KU's EGARC where our recordings took place with the assistance of Paula Li.

6. References

[1] L. E. Armstrong, L. E. and I. C. Ward. 1926. *Handbook of English Intonation*. Cambridge: Heffer.

[2] G. Bailly, and B. Holm. 2005. SFC: A trainable prosodic model, *Speech Communication*, 46:348-364.

[3] Barbosa, P. A. 2011. Panorama of Experimental Prosody Research, in *Proceedings of the VIIth GSCP International Conference – Speech and Corpora*, editors Heliana Mello, Massimo Pettorino and Tommaso Raso, Florence, Italy, 33-42.

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

[5] P. A. Barbosa. 2002. Explaining Cross-Linguistic Rhythmic Variability via a Coupled-Oscillator Model of Rhythm Production, *Proc. Speech Prosody 2002 Conf.* [CD], Aix-en-Provence, 163-166.

[6] P. A. Barbosa, H. Mixdorff and S. Madureira. 2011. Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese, in *Proceedings of Interspeech 2011*, Florence, Italy, 2065-2068.

[7] M. Beckman and J. Pierrehumbert. 1986. *Intonational structure in English and Japanese*. Phonology Yearbook 3 (1986), Cambridge, Cambridge University Press, 255-310.

[8] M. Beckman, J. Hirschberg and S. Shattuck-Hufnagel. 2005. The Original ToBI System and the Evolution of the ToBI Framework, in S.-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press: 9-54.

[9] M. Beckman, M. Díaz-Campos, J. Tevis McGory and T. A. Morgan. 2012. Intonation Across Spanish, in the Tones and Break Indices Framework, in *Probus* 14. Walter de Gruyter, 14, 9-36.

[10] D. Bolinger. 1978. *Intonation Across Languages, in Universals of Human Languages*, vol II Phonology, editor J.H. Greenberg [sic]

[11] D. Brazil. 1975. *Discourse Intonation* [Discourse Analysis Monograph 1]. Birmingham: University of Birmingham, English Language Research.

[12] D. Crystal. 1969. *Prosodic Systems and Intonation in English*, Cambridge, Cambridge University Press.

[13] D. Crystal and D. Davy. 1969. *Investigation English Style*, Bloomington: Indiana University Press.

[14] F. Cummins and R. F. Port. 1998. Rhythmic constraints on stress timing in English, *Journal of Phonetics*, 26(2), 145–171

[15] L. Dilley, S. Shattuck-Hufnagel and M. Ostendorf.

[16] H. Fujisaki. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing, in P. F. MacNeilage [Ed], *The Production of Speech*, 39-55, New York: Springer-Verlag, 1983.

[17] H. Fujisaki and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *Journal of the Acoustical Society of Japan* (E), 5(4): 233–241, 1984.

[18] C. Gussenhoven. 1983. *A Semantic Analysis Nuclear of the Tones of English*. Bloomington, Indiana University Linguistics Club.

[19] D. Jones. 1909. *The Pronunciation of English – A Manual of Phonetics for English Students*, Cambridge, UK: Cambridge University Press.

[20] R. Kingdom. 1958. *The Groundwork of English Stress*, London, Longman.

[21] D. J. Hirst and R. Espesser. 1993. Automatic modelling of fundamental frequency using a quadratic spline function, *TIPA* 15: 71-85, 1993.

[22] D. J. Hirst. 2005. Form and function in the representation of speech prosody, in K. Hirose, D. J. Hirst & Y. Sagisaka (eds) *Quantitative prosody modeling for natural speech description a nd generatio*n (=*Speech Communication* 46 (3-4)), 334-347.

[23] D. R. Ladd. 1980. *The Grammar of Intonation*, Bloomington, Indiana University Press.

[24] J. Moraes. 1998. Intonation in Brazilian Portuguese, in Daniel Hirst and Albert Di Cristo, eds. *Intonation Systems: a survey of twenty languages*, Cambridge, UK: Cambridge University Press, 179-194

[25] J. Moraes. 2008. The Pitch Accents in Braziian Portuguese: Analysis by Synthesis, in *Proceedings of the Fourth COnference on Speech Prosody* 2008. Campinas: RG/CNPq, 389-397.

[26] J. D. O'Connor and G.F. Arnold. 1961. *Intonation of Colloquial English*, London, Longman.

[27] A. R. M. Simões. 2014. Lexical Stress in Brazilian Portuguese in Contrast with Spanish. In Campbell, Nick, Dafydd Gibbon and Daniel Hirst, editors, *Annals of the Speech Prosody Conference # 7*. Dublin, Ireland, Trinity College: 251-255.

[28] A. R. Meireles, J. P. Tozetti1 and R. R. Borges, "Speech rate and rhythmic variation in Brazilian Portuguese," *Proceedings of the 5th International Conference on Speech Prosody,* Chicago, USA, *2010*,

[29] A. R. Meireles and V. de P. Gambarini, "Rhythm Typology of Brazilian Portuguese dialects," *Proceedings of the 6th International Conference on Speech Prosody,* Shanghai, China, *2012*,

[30] H. Palmer. 1922. *English Intonation*, Cambridge: Heffer.

[31] J. Pierrehumbert and J. Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse, in Intentions in Communication. Cambridge, MA: MIT Press.

[32] K. Pike. 1945. *Intonation of English*, Ann Arbor: Michigan University Press.

[33] P. Prieto and J. I. Hualde. 2012. Towards Developing a Standard for Prosodic Annotation, in *Workshop Advancing Prosodic Transcription for Spoken Language Science and Technology*, satellite event to Laboratory Phonology, University of Stuttgart, 13 July 2012.

[34] S. Prom-on and Y. Xu. 2010. Articulatory-Functional Modeling of Speech Prosody: A Review. *Proc. Interspeech 2010*, Makuhari, 46–49, 2010

[35] S. Prom-on, Y. Xu, and B. Thipakorn. 2010. Modeling tone and intonation in Mandarin and English as a process of target approximation, in *J. Acoust. Soc. Am*., 125 (1): 405–424, 2010.

[36] A. K. Syrdal and J. McGory. 2000. Inter-transcriber Reliability of ToBI Prosodic Labelling, in *Proc Int Conf On Spoken Language Processing*, Vol. 3. Beijing, China, 2000, 235-238.

[37] P. Taylor. 2000. Analysis and synthesis of intonation using the tilt model, *J. Acoust. Soc. Am*., 107: 1697-1714, 2000.

[38] TRASP 2013 – *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, an interspeech 2013 satellite event, editors: Brigitte Bigi and Daniel Hirst, August 30, 2013, Aix-en-Provence, France: Laboratoire Parole et Langage,

[39] Y. Xu, and S. Prom-on. 2010. Articulatory-functional modeling of speech prosody: A review, in *Interspeech 2010, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, September 26-30, 2010