Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Acoustic Correlates of Perceived Syllable Prominence in Spanish

Jorge A. Gurlekian[1], *Hansjörg Mixdorff*[2], *Humberto Torres*[1],*Christian Cossio-Mercado*[1], *Diego Evin*[1]

[1]Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Argentina
[2]Department of Computer Sciences and Media, Beuth University Berlin, Germany
jag@fmed.uba.ar, mixdorff@beuth-hochschule.de, hmtorres@conicet.gov.ar
ccossio@dc.uba.ar, diegoevin@gmail.com

Abstract

This paper explores the relationship between perceived syllable prominence and the acoustic properties of a speech utterance. It is aimed at establishing a link between the linguistic meaning of an utterance in terms of sentence modality and focus with its underlying prosodic features. Our acoustic analysis compares traditional parameters modified by focus and sentence mode like fundamental frequency, syllabic durations and intensity against Fujisaki model accent command parameters. Listeners identified narrow focus correctly but only one third of utterances with no focus. Ratings of perceived prominence are moderately correlated with most prosodic parameters. The proportion rate of syllable duration to the underlying accent command duration resulted to be the parameter combination that best correlates to prominence. A simple classifier based on a regression model is presented to detect prominences automatically. This model could explain up to 60% of the observed variance.

Index Terms: prominence, perception, automatic speech recognition, Fujisaki model, F0.

1. Introduction

This work aims to explore linguistic information such as focus and sentence modality and how it can be retrieved from the prosodic features of an utterance. Since a direct link between the acoustics and the meaning of an utterance seems difficult to establish, we decided first to derive syllable prominences from the acoustic signal and then relate these prominence ratings to the focal and sentence mode conditions. Part of this effort is a perceptual evaluation regarding humans ability to retrieve the intended focal condition from isolated utterances. Those acoustic differences which are perceptually salient could be primarely exploited by an automatic speech recognition [1, 2] and understanding system [3]. Such knowledge appears to be also important for other areas of application: in computer-based pronunciation training, speech annotation for data-driven speech synthesis [4], speech summarization [5], speech comprehension for example through improving the syntactic parsing [6], or in audio-visual speech applications by driving gestures for avatars [7].

According to Spanish linguistic rules, all content words are produced with a primary accent located on its stressed syllable [8]. This syllable is to be perceived with a degree of prominence and could be acoustically evidenced by a long duration, high intensity, F0 peaks and a more clearly defined vowel structure and quality [9]. Prosodic prominence follows both linguistic rules and para-linguistic or inferential decisions to convey information. Romance languages formally follow syntactic rules to mark new or enhanced information, pushing the focused constituents to the end of the intonational phrase but naturally the speaker can use those prosodic cues with more or less frequency according to the language- to shift focus to any word [10, 11]. Then based on the speaker's will related to his/her pragmatic communicative intention- any of the acoustic parameters or a combination of them could be further emphasized in the stressed syllable of a focused word [12, 13]. Tonal and phrase accents are candidates to look at. If the emphasis is weak as it occurs in the absence of focus, F0 peaks could be delayed or even precede the stressed vowel position, probably being influenced by their distance to the phrase-juncture boundary tone [14]. When the F0 emphasis is strong, stressed syllables exhibit F0 peaks temporally aligned with them and the phrase is said to have a narrow focus if only one lexical item or part of it is enhanced, or a wide focus if a longer segment is enhanced or even the whole utterance is in focus.

This paper is organized as follows: first, there is a description of the experimental setup and the acoustic analysis of single-phrase utterances produced with varying sentence mode and prosodic no focus and narrow focus. We examine prosodic features derived from Fujisaki accent commands: Amplitude Aa, initial time $T1$ and final time $T2$). Then we measured duration, intensity and the harmonic to noise ratio of the syllables to see how they all correlate with the underlying linguistic information. Second, the perceptual experiments are described in which subjects were asked to determine the sentence mode and focus of the same utterances as well as rate the prominence of each syllable. Third, acoustic analysis results are linked to the outcomes of the perceptual experiment and presented as correlations. Also a regression model is proposed to perform automatic detection of prominences.

In this paper we present results for the Spanish data which are compared to earlier results using German data [15].

2. Experimental Design

2.1. Recordings

We employed nine short sentences created to study on prosody in Argentine Spanish [16] which are part of a multilingual corpus called AMPER (Prosodic Multimedia Atlas for Romance Languages) [17].

Sentences are sintactically restricted to allow a systematic control of variables. The word order in sentences is subject-verb-object and they were created by the full combination of three content words with lexical stress varied systematically in any of three syllables for first critical word in the nominal phrase and the second critical word in the prepositional phrase as shown in Figure 1, for both modalities statements and ques-

 doi: 10.21437/SpeechProsody.2016-138

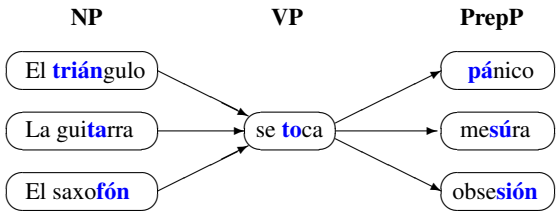

Figure 1: Sentence structure with stressed syllables indicated with bold letters.

tions. Ten college educated subjects - five female and five male - with no phonetic training participated in the recording sessions in a non-acoustically treated room and a dynamic microphone placed near their mouths. The speakers produced three repetitions for each of the nine combinations at a medium speech rate, and the third utterance repetition was chosen. They were told what the purpose of the experiment was and received indications regarding the type of focus that they must produce. Focus types were: 1) no focus, 2) early focus: narrow focus in the first critical word, 3) late focus: narrow focus in the second critical word. Hence 540 utterances were created (9 sentences x 3 focus conditions x 2 modalities x 10 subjects). Recording quality was 44kHz/16bits.

2.2. Acoustic Analysis of Stimuli

Recordings were force-aligned with the Prosody-Lab aligner on the phonetic level [18]. After down sampling to 16kHz automatic syllable segmentations were checked and corrected manually with the ANAGRAF [19] speech analysis system. We subsequently calculated syllable durations based on the segmentations. F0 values were extracted every 10 ms and contours checked and corrected if necessary with ANAGRAF. Besides of evaluating raw F0 contours we approximate each contour with the Fujisaki model by superimposing three components: A base frequency Fb, exponentially decaying phrase components which are the responses to the phrase commands and accent components which are the smoothed responses to the accent commands. We extract the Fujisaki model parameters underlying the F0 contour by applying the method presented in [20] , and evaluate the accent commands, namely their amplitude Aa, and onset time T1 and offset time T2. This way, we are able to measure complete intonational gestures and not only points in time. Fujisaki model parameters were extracted with alpha equal to 2 and beta equal to 20. The alignment of accent commands with syllables was performed automatically based on linguistic information about critical words and their lexically stressed syllables. The extraction method automatically estimates the base frequency for each utterance, and provides an improvement for phrase command insertions. In this method the accent command estimation is linked to stressed syllables in content words. In addition, the final syllable was scanned for high boundary tones also associated with accent commands. Results were checked in the FujiParaEditor [21]. In this way, we obtained a smooth, interpolated model F0 with the accent command amplitudes Aa as a measure of the underlying F0 gesture magnitude and command location given by T1 and T2. Intensity contours were extracted in PRAAT [22] with default settings, and mean intensities in dB, as well as maxima employing parabolic interpolation were determined for each phone. Syllabic mean harmonics-to-noise levels were also calculated within PRAAT with default settings applied.

2.3. Focus and prominence

Seven native listeners participated in a perceptual experiment to evaluate focus condition only in declarative sentences in response to questions related to focus type. Subjects were also asked to indicate the degree of prominence of each syllable in both modalities. The perception test was designed to examine the following research questions:

- Are subjects able to identify the intended sentence modality?
- Can subjects identify the intended focal condition?
- How do subjects rate syllable prominences?

Listeners were familiarized with the task by presenting both text and audio on the screen of different focus examples. For the perceptual experiment, instructions given to subjects were: *"you can listen to each utterance as often as you wish. Then you must decide the modality type between: statement, question or ambiguous. If you choose statement please mark the most suitable question for eliciting that statement. For example, if no focus is present, choose the question Qué pasa? (Whats up?) For narrow focus in the subject please choose the question asking for the instrument, for narrow focus in the object the correct question asks for the way the instrument is played. Finally, adjust the slider levels to indicate the prominence level of each syllable"*. The perception test was performed online and hosted on a server at our laboratory. The experiment was preceded by a verbal explanation of the task. Six examples were presented at the beginning of the experiments. Due to the large number of stimuli the task was rather demanding. Therefore, we recommended that participants should only rate a maximum number of 100 stimuli in one session. The use of an array of sliders for rating prominence was inspired by the works of Eriksson et al. (2001) [23] who employed a similar paradigm in their prominence rating experiments.

3. Results

3.1. Results of the Acoustic Analysis

Figure 2 displays an example of F0 contour decomposition for sentence *"La guitarra se toca con pánico"* (the guitar is played with panic) produced by male speaker SP1 for the focus conditions: (a) no focused statements (b) narrow early *"guitarra"*, (c) narrow late *"pánico"*, (d) no focused questions (e) narrow early-question and (f) narrow late-question. All narrow foci are non-contrastive. Each of the six panels displays from the top to the bottom: the F0 contour (extracted and modeled), and the underlying accent commands. The syllable segmentation is indicated by the dotted vertical lines. Syllable texts are provided in Spanish SAMPA transcription [24]. The natural F0 contour is indicated by circles. The modelled contour is indicated by by xxx signs which approximates the natural contour closely. As can be seen, F0 contours differ clearly for the six conditions. For statements with neutral focus the F0 peak and the corresponding accent command is shifted to the next syllable [Ra], as expected. The narrowly focused words are associated with accent command of high amplitude. The end of the accent command is associated with the F0 peak. Questions are also marked by high levels of accent command amplitude in the focused item. In these examples the beginning of the accent command appears associated with the stressed syllable in the focused word. These examples show also an F0 peak in the intermediate non critical word corresponding to the verbal phrase

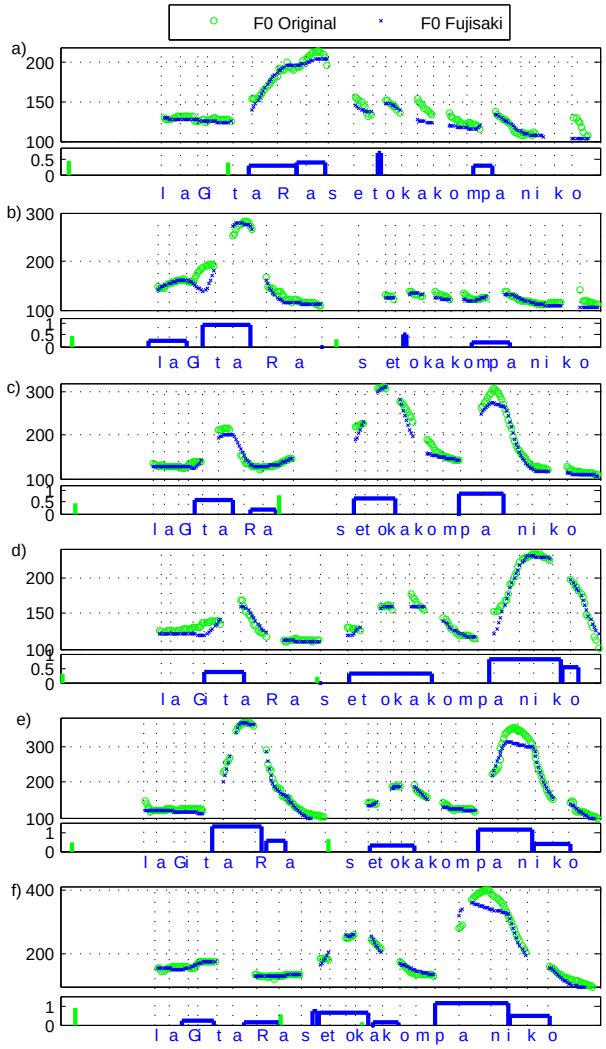

Figure 2: *Results of analysis for male speaker SP1. a) Statements of no focus; b) early focus on* "guitarra"*; c) late focus on* "pánico*; d) questions of no focus; e) early focus on* "guitarra"*; y f) late focus on* "pánico".

[toka]. Figure 3 shows median values and dispersion of accent command amplitude Aa for the three different focus conditions, averaged over the two critical words discriminated by gender and modality. Male speakers show higher Aa for the focused words than the non focused words in statements but this is not the case for the early focus in questions. $T1$ and $T2$ are measured relative to the stressed syllable at the beginning. Mean comparisons between Fujisaki parameters revealed that focus width is significant for the command amplitude, command start $T1$ and the energy of the command defined as $(T2 - T1) * Aa$ and represented in Figure 2 as the area of the accent command, for the first and second critical word (Kruskal-Wallis test of independent samples, $p < 0.001$). Whereas $T2$ is also significant for the second critical word, it is not in the first one. Analysis of contrast distinction for statements was found significant for Aa in the second critical word for early focus vs late focus and for late focus vs no focus (Wilcoxon rank sum test with $p < 0.001$). As shown in Table 1, $T1$ in the first critical word resulted useful to separate early vs late focus and between no focus and narrow

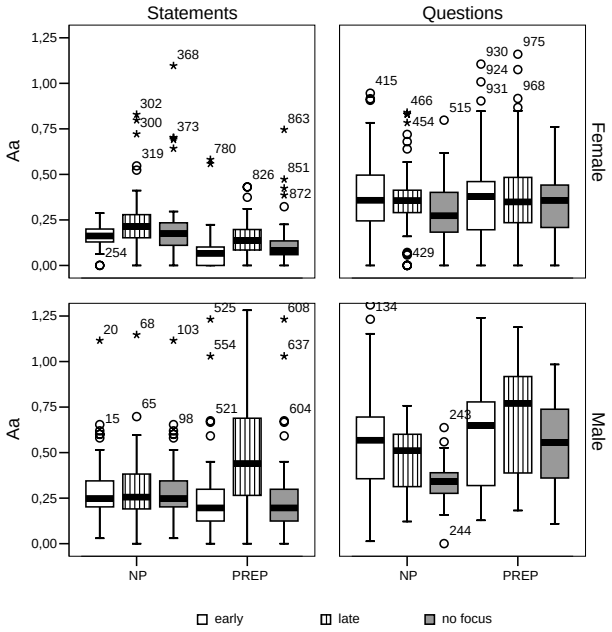

Figure 3: *Box plots of command amplitude Aa, for three focus conditions in the two critical words NP and PrepP and both modality and gender conditions.*

focus. Some combined parameters resulted to be significant to indicate focus position as shown in Table 1: they are command amplitude differences for early and late positions, the energy of the command defined as the amplitude Aa times the command duration $T2 - T1$, and energy differences. Likewise, syllable durations and max intensity in the aforementioned syllables are significantly affected by the focus width (Kruskal-Wallis test of independent samples, $p < .001$).

Table 1: *p-values for the Wilcoxon rank sum test for focus contrasts. Aa, $T1$ and $T2$ are the command parameters. S and Q indicate statements and questions respectively.*

		NP		PrepP	
	Focus	Late	No focus	Late	No Focus
	Mode	S/Q	S/Q	S/Q	S/Q
Aa	Early	0.36/0.06	0.36/0	0/0.17	0.98/0.32
	Late	-/-	0.04/0	-/-	0/0.02
$T1$	Early	0.25/0	0/0	0.94/0.14	0.70/0.03
	Late	-/-	0.07/0.53	-/-	0.78/0.52
$T2$	Early	0.03/0	0/0	0.99/0.24	0.93/0.02
	Late	-/-	0.19/0.93	-/0	0.99/0.30

3.2. Results of perceptual experiments

Listeners identified statements (99.59%) and questions (97.04%) correctly for the majority of the utterances. We examined the responses of our subjects at determining he focus condition of statement utterances. The results based on majority voting are shown in Table 2 . Subjects were asked to select the most suitable question for the statement utterances they listened to and we hoped to identify the perceived focus from their choices. In almost all of these cases the focus location

Table 2: *Confusion matrix showing percentages of correct identification of each focus condition, for statements.*

Intended Focus	No focus	Early	Late
No focus	27.47	1.10	71.43
Early	1.10	90.11	8.79
Late	3.30	1.10	95.60

was identified correctly. This result seems plausible given the results of acoustical analysis presented above. However, only one third of non focused utterances were classified as belonging to this category (27.47%), and even more were considered to be focused on the second critical word (71.43%).

Now we turn to the perceptual prominence ratings of our evaluators. The slider values were mapped onto an integer scale from -5 to +5. Subjects responses were normalized relative to the maximum prominence of each utterance. Responses to each syllable were pooled for every utterance and related to their underlying acoustic features. We first analyze the relationship between the most prominent syllable by majority voting and the command parameter whose value was greater than the rest. See Table 3. The energy of the command appears to be an indicator of prominent syllables better than the other parameters separately. Correlations between prominence and both syllable acoustic features and command parameters are shown in Table 4. Table 4 shows that prominence values are moderately correlated with most parameters. Command amplitude has the highest correlation of 0.56 followed by syllable duration and the energy of the command. Harmonic to noise levels presented a non significant correlation with prominence. We explore various parameter combinations like the F0 maximum within a segment corresponding to the command duration and obtained a Pearson correlation of 0.60 ($p < .001$). A further improvement to 0.65 ($p < .001$) was obtained for the product of F0 delimited by the command duration times the syllable duration. Maximum correlation of 0.72 ($p < .001$) was obtained for the proportion of syllable duration to command duration.

Table 3: *Majority votes of prominence judgments vs. command parameters shown as relationship in percentages.*

Parameter	Aa		$T2-T1$		$Aa \cdot (T2-T1)$	
Gender	M	F	M	F	M	F
Statements	69	71	68	65	83	80
Questions	76	77	61	81	83	81
Average	73	74	64	73	83	80

Table 4: *Correlations between perceptual prominence and selected prosodic features of syllables, as well as correlations between these features. Pearson's r, two-sided significance value. Number of instances 5379. Significance levels: $p < .001$.*

Feature	F0	Dur.	Int.	Aa	$T2-T1$	E
Promin.	0.32	0.55	0.31	0.56	0.44	0.50
F0		0.24	0.37	0.35	0.25	0.38
Syll. dur.			0.17	0.45	0.33	0.47
Intensity				0.22	0.13	0.20
Aa					0.62	0.90
$T2-T1$						0.71

Table 5: *Regression model for predicting perceptual syllable prominence based on the factors listed in the left column.*

Factor	B	Std. error	T	sign
(constant)	-3.323	0.2582	-12.867	0.0001
F0	0.002	0.0002	7.447	0.0001
Syll. duration	0.007	0.0002	32.061	0.0001
Intensity	0.045	0.0034	13.413	0.0001
Aa	2.888	0.1376	20.988	0.0001
$T2-T1$	2.110	0.1534	13.760	0.0001
$Aa \cdot (T2-T1)$	-6.616	0.5800	-11.406	0.0001

Further regression models were obtained for the above mentioned parameter combinations as shown in Table 5. When we add F0 maximum delimited by command duration to the parameters indicated in Table 4, and the command energy and the proportion of syllable duration to command duration we were able tod explain 57.6% of the variance.

4. Discussion and Conclusions

We presented results from a production and perception study aiming at determining the effect of focus and sentence mode on several prosodic parameters, as well as the connection between perceived focus and syllable prominence with these parameters. The ultimate aim is to determine this kind of linguistic information in automatic speech recognition and enrich the word hypothesis.

Our acoustic results are in line with the study made for German [15] with respect to effects of F0 range expansion, increased duration and intensity in focused items and the reverse for the de-focused ones. Main differences are for questions where Argentine Spanish utterances show peaks in focused words followed by an F0 fall [16, 25, 26] which is not observed for the majority of all other Spanish varieties. For German F0 remains high from the last focussed word to the end of the utterance. Non focused utterances are clearly distinguished from those with narrow or late focus on one of the two critical words. However, this result was not matched by perceptual outcomes in both languages. Utterances with intended no focus were only identified in about one third of cases. The non focused question was generic whereas the others were related to one of the critical words. This idea is supported by the observation that decisions towards *"no focus"* in the focused cases were relatively few and almost exclusively occur in *"late focus"* cases. This latter result points out the fact that according to linguistic rules non focused utterances always imply a prominence on the last accentable item [27, 28]. In our case it is the second critical word, making the choice *"late focus"* a plausible one. We also found that prominence ratings can be fairly well predicted based on the combination of both command properties and syllable duration. Although our perception results somewhat question the ability of human listeners to reliably detect focus, if ASR were able to enhance the word string with prominence ratings, this would already represent a step forward.

5. Acknowledgements

This work was funded through Argentine-German Bilateral Project Mincyt-BMBF AL12.

6. References

[1] M. Hasegawa-Johnson and S. B. k. Chen, "Experiments in landmark-based speech recognition," in *Proc. of Sound to Sense: Workshop in Honor of Kenneth N. Stevens*, June 2004.

[2] S.-H. Chen, J.-H. Yang, C.-Y. Chiang, M.-C. Liu, and Y.-R. Wang, "A new prosody-assisted mandarin asr system," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1669–1684, 2012.

[3] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The use of prosody in the linguistic components of a speech understanding system," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 5, pp. 519–532, 2000.

[4] A. Windmann, P. Wagner, F. Tamburini, D. Arnold, and C. Oertel, "Automatic prominence annotation of a german speech synthesis corpus: towards prominence-based prosody generation for unit selection synthesis," in *Proceedings of the 7th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2010.

[5] D. Z. Hakkani-Tür, G. Tür, A. Stolcke, and E. Shriberg, "Combining words and prosody for information extraction from speech," in *EUROSPEECH*, Budapest, Hungary, September 1999, pp. 1991–1994.

[6] L. Frazier, K. Carlson, and C. Clifton, "Prosodic phrasing is central to language comprehension," *Trends in Cognitive Sciences*, vol. 10, no. 6, pp. 244–249, 2006.

[7] S. Al Moubayed, J. Beskow, B. Granström, and D. House, "Audiovisual prosody: Perception, detection, and synthesis of prominence." in *COST 2102 Training School*. Springer, 2010, pp. 55–71.

[8] T. L. Face, "Efectos segmentales del acento en español," *Boletín de lingüística*, no. 14, pp. 18–32, 1998.

[9] D. B. Fry, "Experiments in the perception of stress," *Language and speech*, vol. 1, no. 2, pp. 126–152, 1958.

[10] L. O. Labastía, "Prosodic prominence in argentinian spanish," *Journal of Pragmatics*, vol. 38, no. 10, pp. 1677–1705, 2006, special Issue: Prosody and Pragmatics.

[11] J. D. Luis, "La focalización prosódica: funcionalidad en los niveles lingüistico y pragmático," *Estudios de fonética experimental*, vol. 17, pp. 106–138, 2008.

[12] G. Toledo, "Señales prosódicas del foco," *Revista Argentina de lingüística*, vol. 5, no. 1-2, pp. 205–230, 1989.

[13] T. Face, "Prosodic manifestations of focus in spanish," *Southwest Journal of Linguistics*, vol. 19, no. 1, pp. 45–62, 2000.

[14] J. Gurlekian, H. Mixdorff, D. Evin, H. Torres, and H. Pfitzinger, "Alignment of F0 model parameters with final and non-final accents," in *Proc. of 5th International Conference on Speech Prosody 2010*, Chicago, Illinios, USA, May 2010.

[15] H. Mixdorff, C. Cossio-Mercado, A. Hönemann, J. Gurlekian, D. Evin, and H. Torres, "Acoustic correlates of perceived syllable prominence in German," in *Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany, September 2015, pp. 51–55.

[16] J. Gurlekian and G. Toledo, "Datos preliminares del amper-argentina: las oraciones declarativas e interrogativas absolutas sin expansión," *Language Design, J. of Theoretical and Experimental Linguistics Special Issue: Experimental Prosody*, vol. 2, pp. 213–220, 2008.

[17] M. Contini, J.-P. Lai, A. Romano, S. Roullet, L. d. C. Moutinho, R. L. Coimbra, U. P. Bendiha, and S. S. Ruivo, "Un projet d'atlas multimédia prosodique de l'espace roman," in *Speech Prosody 2002, International Conference*, Aix-en-Provence, France, April 2002, pp. 227–230.

[18] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, vol. 39, no. 3, pp. 192–193, 2011.

[19] J. A. Gurlekian, *El laboratorio de audición y habla del LIS*, M. Guirao, Ed. Buenos Aires: Dunken, 1997.

[20] H. Torres and J. Gurlekian, "Novel estimation method for the superpositional intonation model," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 24, no. 1, pp. 151–160, 2016.

[21] H. Mixdorff, "Fujiparaeditor: http://www.tfh-berlin.de/∼mixdorff/thesis/fujisaki.html," 2009.

[22] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[23] A. Eriksson, G. C. Thunberg, and H. Traunmller, "Syllable prominence: a matter of vocal effort, phonetic distinct-ness and top-down processing." in *INTERSPEECH*, Aalborg, Denmark, September 2001, pp. 399–402.

[24] J. A. Gurlekian, L. Colantoni, and H. M. Torres, "El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados," *Fonoaudiológica*, vol. 47, no. 3, pp. 58–70, 2001.

[25] S. A. Lee, "Absolute interrogative intonation patterns in buenos aires spanish," Ph.D. dissertation, The Ohio State University, 2010.

[26] G. Toledo and J. Gurlekian, "Interrogativas absolutas con expansión, el caso marcado," *Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile*, no. 23, pp. 41–68, 2011.

[27] M. L. Zubizarreta, "Las funciones informativas: tema y foco," in *Gramática descriptiva de la lengua española*. Espasa Calpe, 1999, pp. 4215–4244.

[28] J. P. Barjam, "The intonational phonology of porteño spanish," Master's thesis, University of California, Los Angeles, USA, 2004.