Alignment of Intonational Events in German and Brazilian Portuguese – a Quantitative Study

Hansjörg Mixdorff [1], *Plínio A. Barbosa*[2]

[1] Dept. of Computer Science and Media, Beuth University of Applied Sciences, Berlin, Germany
[2]Speech Prosody Studies Group, Dept. of Linguistics, State Univ. of Campinas, Brazil
mixdorff@beuth-hochschule.de, pabarbosa.unicampbr@gmail.com

Abstract

This study compares the intonational realizations of reading style as well as story-telling in German and Brazilian Portuguese (BP). By applying the Fujisaki model we can examine the frequencies and alignments of tonal transition, the so-called tone switches. Analysis of accent commands from the different corpora shows that in German 75% of commands can be associated with automatically generated stress labels. In comparison, manually assigned prominence labels can account for 80% of commands. This latter figure is much lower for BP (61%). This suggests that in BP prominence is signaled by *F0* cues to a lesser degree. In both languages story-telling is associated with fewer accented syllables, a higher rate of rising tone switches on stressed syllables and a broader distribution of accent command amplitudes than reading style.
Index Terms: intonation modeling, speaking styles, cross-linguistic prosody

1. Introduction

The segmental anchoring of F0 contours has been an important research topic for many years and it has been debated to what extent alignment choices are mainly results of phonetic realization or conscious phonological choices speakers of a certain language make [1]. For the case of German Isačenko and Schädlich [2] and Stock and Zacharias [3] describe a given *F0* contour as a sequence of communicatively motivated tone switches, major transitions of the *F0* contour aligned with accented syllables.

In order to quantify the interval and timing of the tone switches with respect to the underlying syllables, we adopt the Fujisaki model [3] which reproduces *F0* from three components: Base frequency *Fb*, phrase component and accent component. Here we are mainly concerned with the alignment of the accent component, the response of the model to step-wise accent commands, defined by on- and offset times *T1* (*F0* rise) and *T2* (*F0* fall), and amplitude *Aa*. In earlier contributions by the first author and his co-workers we investigated the F0 contour anchoring at syllables of early, middle and late peaks [5].

Hence by statistically evaluating the timing and amplitude of accent commands as well as the frequencies of tone switch types we can describe the characteristics of F0 anchoring observed inside a certain corpus.

In the current study we apply this methodology to reading style and story-telling speech in German as well as Brazilian Portuguese (BP). We aim to find the differences of alignment choices with respect to tone switch types, as well as the precise anchoring with the underlying syllabic unit as functions of language and speaking style.

Fujisaki model parameters are typically estimated from the observed *F0* contour without applying linguistic knowledge [6] and the resulting accent commands are automatically aligned to the syllabic grid by associating each onset or offset of a command to the syllable in which it occurs. This syllable, however, is not necessarily the accent-bearing one. Since the tone switches are properties of the accented syllables, command associations must therefore be verified and corrected manually. Only accent commands that can be justified by accented syllables or boundary tones are preserved, others deleted and the optimization is rerun using the modified number of commands. In the case of large corpora this post-processing is prohibitively time-consuming.

Therefore, in the current study we examine to what extent the automatic alignment can be guided by pre-existing information on lexically stressed syllables (1). We compare this option to an alignment based on manually marked perceptually prominent syllables (2). In approach (1) we process the utterance texts with a TTS front-end generating hypotheses on stressed syllables (only German). In approach (2), phonetically trained human labellers mark the words they perceive as most prominent in an utterance. Based on this information we compare German and BP with respect to their intonational characteristics. This study is part of a joint project in which we aimed to compare the benefits of the *qTA* model [7] and the Fujisaki model.

2. Speech Material and Method of Analysis

Two female and seven male speakers (German) and two female and four male speakers (BP) – students of Linguistics or Computer Science - read a 1,500-word text about the pastries “Pastéis de Belém” (reading, RE) in their respective languages [7]. Later, subjects retold the story (story-telling, ST). F0 values were extracted using the standard method in *Praat* [8] at a step size of 10 ms and inspected for errors. The F0 tracks were subsequently decomposed using the standard automatic method [6] and if necessary corrected using the *FujiParaEditor* [9]. For the German data, a TTS front-end was applied to the texts of the utterances predicting stressed and unstressed syllables [10]. For both German and BP two and three labellers, respectively, marked perceptually prominent syllables. In the following description, we refer to syllables that are marked as lexically stressed or perceptually prominent as “marked”.

A computer program was developed which associates the accent commands from the automatic estimation with marked syllables. In a first search each syllable which exhibited onsets or offsets of accent commands was labeled accordingly. Then it was checked whether the current syllable was of the marked type and several alignment options evaluated. Depending on the situation found, the tone switch associated with the syllable is classified as either rising or falling. The search takes into account the current marked syllable and its immediate left and right neighbours. The following are the most frequent cases, as the results section will show (examples are indicated with reference to Figure 2, respective syllables in

bold face): **(1)** A command starts inside the marked syllable and ends in one of the following ones - rising tone switch (Figure 2(1): "**Ma**nuel"). **(2)** An accent command starts and ends inside the marked syllable – rising/falling tone switch (Figure 2(4): "Manu**el**"). **(3)** A command begins in one of the preceeding syllables and ends inside the marked syllable - falling tone switch (Figure 2(1): "**Klos**ter"). **(4)** A command ends in the marked syllable and another one begins – falling/rising tone switch (Figure 2(3): "**qua**se"). The analyses shown here concern excerpts of 150 to 200 words in each language and style and we present a first global and mostly statistical description of results.

3. Results

Figure 2 shows examples of Fujisaki model-based analysis from both languages and speaking styles. Each of the four panels displays from the top to the bottom: the speech waveform, the *F0* contour (extracted and modeled), as well as the underlying phrase and tone commands. The syllable boundaries are indicated by the dotted vertical lines. The texts and English translations of all utterances are given in the caption. Syllables marked as prominent are set in bold face.

As can be seen, all syllables marked as prominent are associated with accent commands. For instance, in example (1), the first syllable of the name "**Ma**nuel" exhibits the onset of an accent command, hence is associated with a rising tone switch on this syllable. In contrast, the offset of an accent command occurs in the first syllable of the word "**Klos**ter", yielding a falling tone switch. However, sometimes the accent command does not begin or end inside a prominent syllable. In example (3), the rising tone switch on the third (prominent) syllable of "Manu**el**" begins slightly before the onset of that syllable. Furthermore, the falling tone switch associated with the word "**a**no" already occurs in the preceding syllable "um". By considering syllables adjacent to prominent syllables the most plausible association can be made.

German. The TTS front-end predicts 26.2% of syllables as stressed, 84% of which had been labeled as prominent (RE 88.4%, ST 78.8%). In RE 30.4% of syllables are labeled as prominent, in ST only 22.1%. Of the total of 1614 accent commands from the analysis, 75% were associated with lexically stressed syllables predicted by the TTS-frontend, and 80% with perceptually labeled ones. Hence the perceptual labels permit slightly better alignments than the automatic lexical stress assignment. Of the remaining accent commands 17% were found in pre-boundary syllables that do not carry lexical stress, but are associated with high boundary tones.

Figure 1 displays histograms of accent command amplitudes *Aa* for RE (top) and ST (bottom). Although in both cases the mean of *Aa* is .31, the standard deviation of .20 is much larger in ST (N=790) than in RE (s=.14, N=824). The strong left skew of the distribution indicates that in ST more accents are weaker than in reading style, but some exhibit rather high amplitudes. Syllables on the average are shorter in RE (mean=170 ms), but also more regularly spaced (s.d.=77 ms) than in ST (mean=196 ms, s.d.=115 ms). Table 1 displays the alignment of prominent syllables with accent command onsets (rises) and offsets (falls) in %. "None" denotes no command close to the syllable. Syllables located during an accent command, but far from its onset and offset are indicated by "command across". There are more rises in ST than RE. This is typical for story-telling because rising tone switches raise attention and establish contact with the listeners. This explains why the distribution of *Aa* for ST in Figure 1 (bottom) extends towards much higher values than that for RE. Facts are often connected in a long string of enumerations which demand rising tone switches. Good examples of these high continuation rises can be witnessed on the word "**Klos**ter" in example (2) and on the words "Manu**el** and "Je**rô**nimos" in example (4). In contrast, in reading style, full stops are usually associated with falling tone switches (see word "**Klos**ter" in example (1) and "**a**no" in example (3)).

Brazilian Portuguese. In RE 19.6% of syllables are labeled as prominent, in ST 17.9%. Of the total of 1125 accent commands from the analysis, only 61% were associated with perceptually prominent syllables.

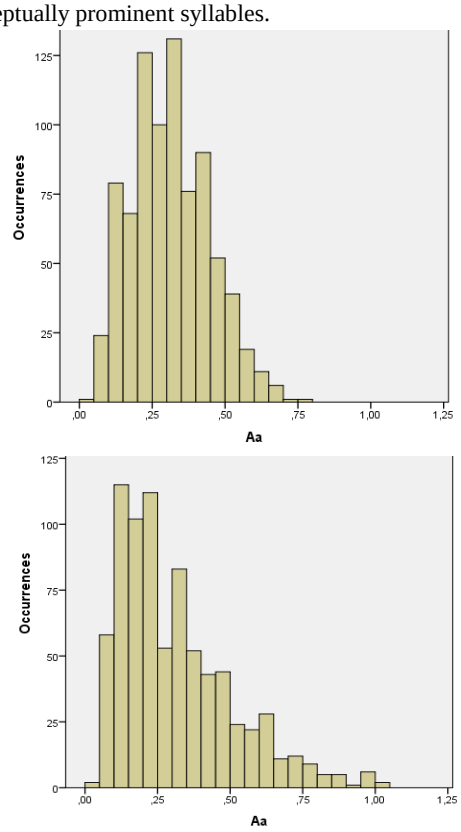

Figure 1:*Histograms of accent command amplitude Aa for German, top RE, bottom ST.*

Table 1: *Alignment of prominent syllables with accent commands onsets (rises) and offsets (falls) in %. "None" denotes no command close to the syllable. Syllables located during an accent command, but far from its onset and offset are indicated by "command across". Numbers (1)-(4) refer to alignments explained in Section 2.*

corpus	(1)rising	(2) rising-falling	(3) falling	(4) falling-rising	fall in preceding	rise in following	other	command across	none
German RE	41.2	14.4	12.6	1.5	1.7	1.4	15.5	2.3	9.4
German ST	44.3	13.2	8.7	2.2	1.2	1.4	15.2	3.3	10.5
BP RE	35.2	9.2	18.1	8.7	5.5	1.7	13.6	5.0	3.0
BP ST	33.4	9.2	11.5	17.4	3.0	1.3	15.3	6.6	2.3

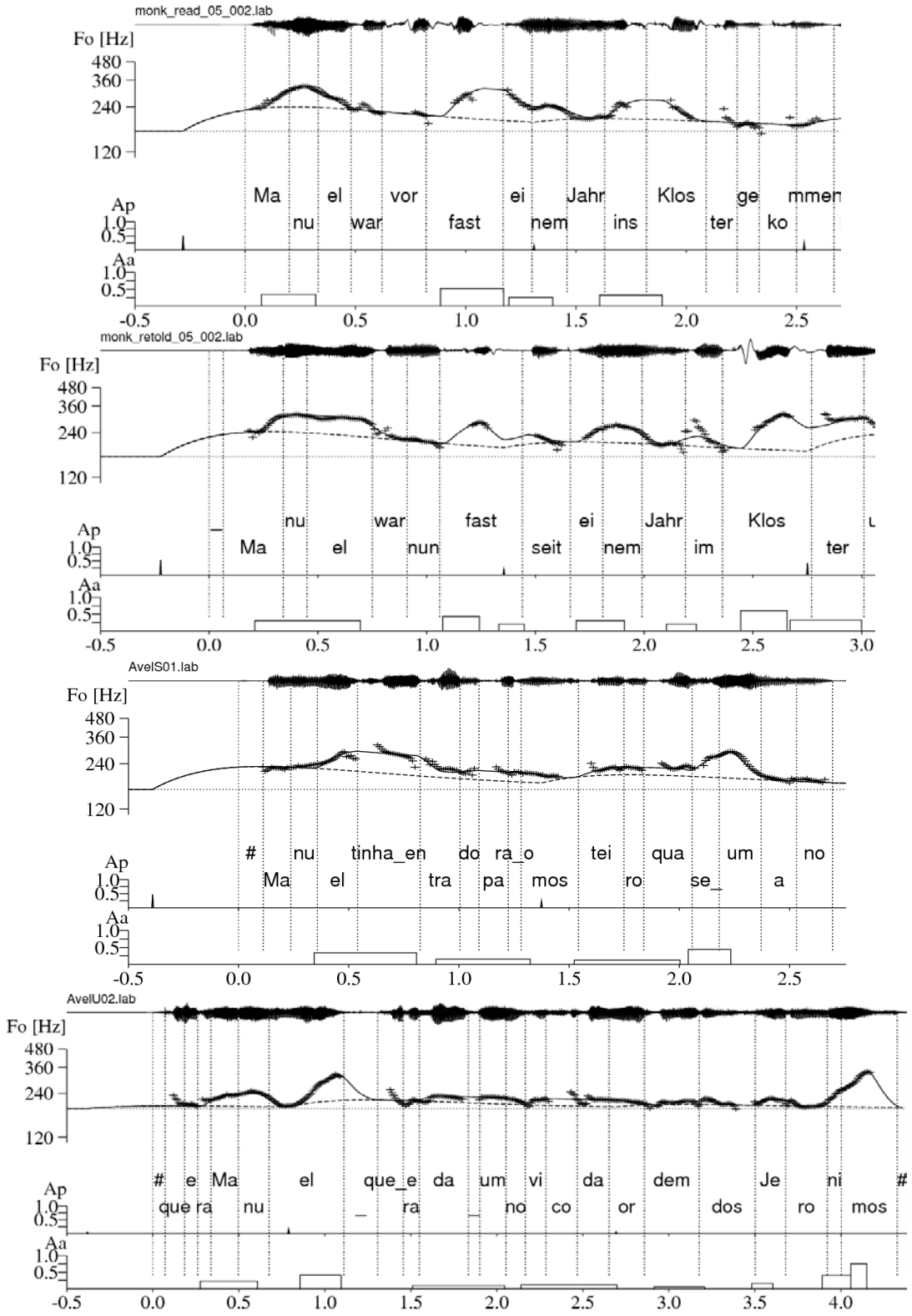

Figure 2: *Four examples of analysis. From top to bottom: (1) German female speaker 5, RE, "**Ma**nuel war vor **fast** einem Jahr ins **Klo**ster gekommen"-"Manuel had come to the monastery almost one year ago." (2) German female speaker 5, ST: "**Ma**nuel war nun **fast** seit **ei**nem Jahr im **Klo**ster…"-"Manuel had been in the monastery for almost one year..."; (3) Brazilian female speaker 1, RE: "Manu**el** tinha entrado para o mos**tei**ro há quase um **a**no." –see (1); (4) Brazilian speaker 1, ST: "…que era Manu**el**, que era da… um no**vi**ço da **or**dem dos Je**rô**nimos." –"who was Manuel, who was of... a novice of the Order of St. Jerome…"*

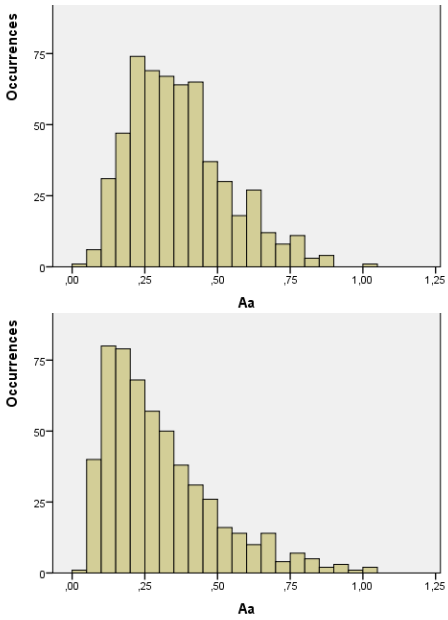

Figure 3:*Histograms of accent command amplitude Aa for BP, top RE, bottom ST.*

However, 97% of these prominent syllables were aligned with accent commands. Figure 3 displays histograms of accent command amplitudes *Aa* for RE (top) and ST (bottom). It shows significantly larger *F0* ranges in RE (mean of *Aa*=.37, N=577) than in ST (mean of *Aa*=.30, N=548) and also compared to German (mean=.31). Similar to German, there is a larger proportion of smaller tone switches in ST (with *Aa* < .25) than in RE and a few larger ones (see also Figure 2, example (4)). Table 1 suggests less preference for rising tone switches than in German. If we look at the frequency of accent commands in both languages and speaking styles, the figures are 1.39 and 1.20 per second for German in RE and ST, respectively, and 1.52 and 1.32 per second for BP. This, together with the ratios of prominent syllables, suggests that subjects accent more often when they read. We calculated the alignment of simple rising and falling tone switches with respect to the segmental onset of the current syllable (cases in columns 2 and 3 of Table 1). In German rises occur at 80/58 ms (mean/s.d.) into the syllable and 84/60 ms for BP, falls at 116/78ms for German and 131/88 ms for BP. These values suggest that there is no significant difference regarding the close alignment of *F0* rises and falls with the prominent syllable in both languages. Only the distributions of alignment situations differ. Similar to German, syllables on the average are shorter in RE (mean=155 ms) but also more regularly spaced (s.d.=71 ms) than in ST (mean/s.d.=191/123 ms).

4. Discussion and Conclusions

In this still preliminary study we compared the frequencies and alignments of tone switches underlying *F0* contours in German and BP for two speaking styles, reading and story-telling. For German we examined whether the alignment of tone switches can be guided by automatically generated lexical stress labels as compared to manually assigned prominence labels. Our results show that for German only 75% of accent commands can be accounted for by lexical stresses. If the alignment is based on prominent syllables, this figure rises to 80%. This means that the automatic alignment invariably misses smaller accent commands for either method. The ratio yielded with the TTS front-end, however, seems high enough to perform statistical analysis of the frequencies, timings and intervals of tone switches, provided the corpus is large enough. For BP, the percentage of accent commands associated with prominent syllables is much lower than for German. Despite this fact, the comparable frequencies of accent commands indicate comparable movement of *F0* in the two languages. This suggests that prominence labels were assigned more sparsely in BP and hence *F0* movement does not contribute as much to the percept of prominence as in German (see [11], for instance). In fact, many of the accent commands extracted from BP *F0* contours are rather long, have small amplitude and extend over many syllables (compare, for instance, Figure 2, (3) and (4)) Overall the mean duration of commands on the BP corpus of 330 ms is longer than for German (280 ms). The two speaking styles differ, inter alia, with respect to the accent command amplitude distributions which are more spread and with a skew towards smaller values for story-telling. In the future we will examine in finer detail the agreement between speakers, as well as the differences between the two languages with respect to the way content words in the reading text are assigned prominence.

5. Acknowledgements

This project was supported by grant 490726/2008-9 from CNPq and 444 BRA 113 59 0-1 from DFG. The second author also acknowledges the CNPq grant 300371/2008-0. The BP corpus was recorded in the context of the FCT project PTDC/PLP/72404/2006, for which INESC-ID Lisboa had support from the “Quadro Comunitário de Apoio III”.

6. References

[1] Atterer, M. and D. R. Ladd, “On the phonetics and phonology of segmental anchoring of F0: evidence from German”. *J. of Phonetics 32*, 177-197, 2004.

[2] Isačenko, A.V., Schädlich, H.J., “Untersuchungen über die deutsche Satzintonation”, Akademie-Verlag, Berlin, 1964.

[3] Stock E., Zacharias, C., “Deutsche Satzintonation“, VEB Verlag Enzyklopädie, Leipzig, 1982.

[4] Fujisaki, H. and Hirose, K. “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, J. of the Acoustical Society of Japan (E) 5(4), 233-241, 1984.

[5] Mixdorff, H. Pfitzinger, H., “A Quantitative Study of F0 Peak Alignment and Sentence Modality”, in Proceedings of Interspeech 2009, Brighton, England, 2009.

[6] Mixdorff, H., “A novel approach to the fully automatic extraction of Fujisaki model parameters”, in Proc. of ICASSP, Istanbul, 3:1281–1284, 2000.

[7] Barbosa, P., Mixdorff, H. and Madureira, S., “Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese”, in Proceedings of Interspeech 2011, Florence, Italy.

[8] Boersma, Paul. “Praat, a system for doing phonetics by computer”. *Glot International* **5:9/10**, 341-345, 2001.

[9] Mixdorff, H. (1/10/2009). *FujiParaEditor*, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html

[10] Hilbert, A., and Mixdorff, H.,“Weiterentwicklung eines Sprachsynthesesystems”, in G. Görlitz [Ed.], Nachhaltige Forschung in Wachstumsbereichen Band I, Logos Verlag, Berlin, 35-42, 2011.

[11] Barbosa, P.A., Silva, W. da, “A New Methodology for Comparing Speech Rhythm Structure between Utterances: Beyond Typological Approaches”, submitted to Propor 2012 - International Conference on Computational Processing of Portuguese, Coimbra, Portugal.