ON THE ALIGNMENT OF PROSODIC EVENTS

Hansjörg Mixdorff

Department of Computer Science and Media, Beuth University of Applied Sciences, Berlin, Germany

mixdorff@beuth-hochschule.de

Abstract: The current study examines the relationship between intonational gestures as given by the accent commands of the Fujisaki model and the syllabic grid on the example of spontaneous American English from the Buckeye Corpus. As an initial step the data were labelled according to American English ToBI conventions. Intensity contours were extracted from the band-filtered speech signal and modelled using a second-order linear model like the accent control mechanism of the Fujisaki model. It was observed that certain accent types such as L* occasionally required the use of accent commands with negative polarity. The timing properties of the underlying accent commands correspond to the type of accent label, i.e., for instance, early for H*L and late for L*H. Accents in the vicinity of a boundary exhibited slightly higher accent command amplitudes whereas the highest syllable command amplitudes were observed for phrase-initial and medial accents. With respect to the alignment of accent commands, the onsets of syllable commands did not prove to be more precise anchoring points than the segmental boundaries of the syllable.

1 Introduction

On the occasion of Prof. Fujisaki's 80th birthday I present a study from 2008 on which I gave a talk at Acoustics `08 in Paris, but which for several reasons never made it into any kind of proceedings. Neither was a written version mandatory at that conference, nor did we ever find out what happened to the full papers that were actually submitted, like my colleague Hartmut Pfitzinger's. But that's a different story.

Over more than 18 years Prof. Fujisaki has been my mentor and I remember very well the time I worked at his lab at Science University of Tokyo from October '93 till March '95. Although our relationship did not start out in utter harmony, over the years it developed into a strong, respectful, even amicable one. Without him I would not be where I stand today. There are few scholars of my generation who could boast a background as wide as Prof. Fujisaki's. Maybe this is only logical as the field of speech processing has grown immensely over the decades of his professional life. And even nearing the age of 80 I still often see him engage in lively discussions, for instance, at Speech Prosody 2010, providing critical, but mostly constructive comments. Some mistake his passion for the topic as arrogance. And I also know that many people associate Prof. Fujisaki foremost with the command-response model which he developed [1] and which I have been using in much of my work. Although this may be one of his most original contributions and one which he is obviously endeared with, I am tempted to say that it is just a part of what he contributed to the speech community as a scholar, an organizer, and a caring teacher.

At the same time when I became familiar with the Fujisaki model I read the works of Isačenko and Schädlich [2] and Stock and Zacharias [3], which influenced my thinking considerably. According to their a approach a given *F0* contour is mainly described as a sequence of communicatively motivated tone switches, major transitions of the *F0* contour aligned with accented syllables. Tone switches can be thought of as the phonetic realization of phonologically distinct intonational elements, the so-called intonemes. In the original formulation by Stock, depending on their communicative function, three classes of intonemes are distinguished, namely the N↑ intoneme (non-terminal intoneme, signalling incompleteness and continuation, rising tone switch), I↓ intoneme (information intoneme at declarative-final accents, falling tone switch, conveying information), and the C↑ intoneme (contact intoneme associated, for instance, with question-final accents, rising tone switch, establishing contact). Hence intonemes in the original sense mainly distinguish sentence modality, although there exists a variant of the I↓ intoneme, I(E)↓ which denotes emphatic accentuation and occurs in contrastive, narrowly focused environments. Intonemes for reading style speech are predictable by applying a set of phonological rules to a string of text as to word accentability and accent group formation.

Based on this concept, Mixdorff and Jokisch [4] developed a model of German prosody anchoring prosodic features such as F0, duration, and intensity to the syllable as a basic unit of speech rhythm. In order to quantify the interval and timing of the tone switches with respect to the syllabic grid, the framework adopts the Fujisaki model for parameterizing *F0* contours [1]. The Fujisaki model reproduces a given *F0* contour by superimposing three components: A speaker-individual base frequency *Fb*, a phrase component and an accent component. The phrase component results from impulse responses to impulse-wise phrase commands associated with prosodic breaks. Phrase commands are described by their onset time *T0*, magnitude *Ap* and time constant *alpha*. The accent component results from step-wise accent commands associated with accented syllables. Accent commands are described by on- and offset times *T1* and *T2*, amplitude *Aa* and time constant *beta*. Phrase and accent command timings are related to syllable onsets and offset as illustrated in Figure 1

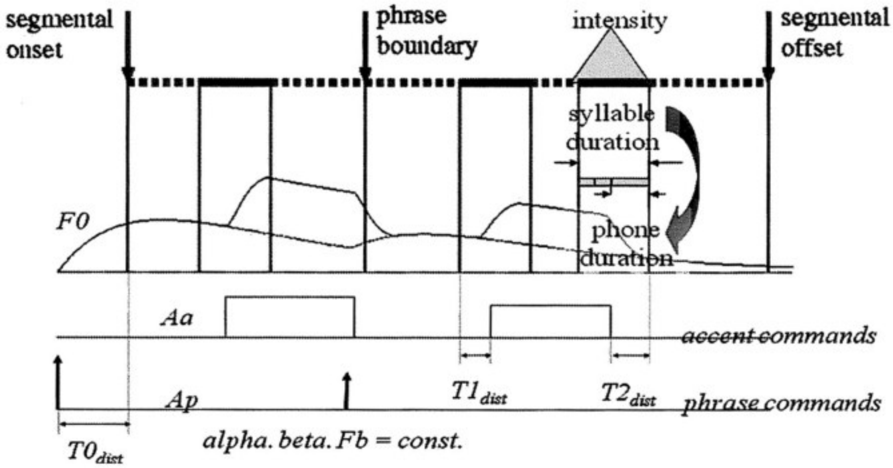

Figure 1 - An illustration of how Fujisaki model parameters are anchored to the syllabic grid. The timing of accent and phrase commands is related to the segmental onsets and offsets of syllables.

In a perception study [5] employing synthetic stimuli of identical wording but varying *F0* contours created with the Fujisaki model it was shown that information intonemes are characterized by an accent command ending before or early in the accented syllable, creating a falling contour. N↑ intonemes were connected with rising tone switches to the mid-range of the subject connected with an accent command beginning early in the accented syllable and plateau-like continuation up to the phrase boundary, whereas C↑ intonemes required *F0* transitions to span a total interval of more than 10 semitones and generally starting later in the accented syllable, although the *F0* interval was a more important factor than the precise alignment.

Mixdorff and Fujisaki [6] compared German ToBI labels with Fujisaki model parameters on a corpus of news reading. They found that tone labels were strongly correlated with accent commands, and the type of label (typically H*L and L*H) was clearly reflected by the onset and offset times of these accent commands. These main label types once again correspond to the I↓- and N↑ intonemes in Stock's formulation, respectively. Pfitzinger and Mixdorff compared PROLAB labels of the Kiel intonation model on a corpus of spontaneous speech [7] and showed a close relationship between early, medial and late peaks and the timing of the underlying accent commands.

The current study is intended to investigate the alignment between intonational and articulatory gestures by relating accent commands of the Fujisaki model to so-called syllable commands modelling the intensity contour of the speech signal. In earlier works accent command onsets and offsets were either aligned with syllabic boundaries [4], or with respect to the vowel nucleus [7]. In the current study onsets and offsets of syllable commands derived from the intensity contour are examined as alternative anchoring points. The model is applied to American English spontaneous speech and used to examine the relationship between ToBI labels [8] and accent commands.

2 Speech Material and Method of Analysis

The speech material consists of a subcorpus from the *Buckeye Corpus of Conversational Speech* [9] of approximately 9 minutes by a single male talker, containing a total of 1729 syllables. The corpus contains annotations on the phone and word levels. These were augmented by syllable, phrase and topic levels.

The audio level for the interviewer was very low making it hard to follow the interaction with the interviewee. Therefore a dynamics-adjusted version was created which allowed to text-annotate the turns of the interviewer. The perceptually salient syllables were labelled with American ToBI tone labels and phrase breaks using break indices [10].

F0 values were extracted at a step of 10ms using the *PRAAT* default pitch extraction settings [11]. Then Fujisaki model parameters were estimated [12] (*Fb*=95Hz, *alpha*=2/s, *beta*=20/s) and manually corrected in the *FujiParaEditor*[13]. Occasionally, accent commands of negative polarity were used to model F0 at low L* accent syllables where an obvious active lowering occurred.

As a concomitant of the articulatory gesture we calculated log intensity contours of a band pass (300-1500 Hz) filtered version of the speech signal, also using *PRAAT*. These intensity contours were then modelled using box-shaped syllables commands similar to accent commands with a time constant beta of 80/s and a floor value of 40dB. The hill-climb search for optimizing the syllable commands was performed in a similar fashion to that for accent commands in [12].

3 Results of Analysis

Figure 2 displays a result of F0 contour analysis of the sentence "They stopped federally funding OSU". The panel contains from the top to the bottom: The speech wave form, the *F0* contour (+signs: extracted, solid line: Fujisaki model-based), the syllabic text, the underlying phrase and accent commands. The first part of the sentence exhibits hesitation "uh they uh…" where the F0 contour essentially follows the phrase component. As can be seen, the accent on "OSU" which was assigned the tone label H*L is modelled using a pair of accent commands with positive/negative polarities. Figure 3 displays an example of intensity modeling of the sentence "Well, it's the government and regulations..." From the top to the bottom: The speech waveform, syllabic text, the intensity contour (+ extracted and - modeled) and underlying syllable commands. The word "well" is actually pronounced "we-ell" and therefore modeled using two syllables commands. Table 1 displays the correspondence between ToBI labels and accent commands. 16% of syllables bearing tone labels were not associated with accent commands.

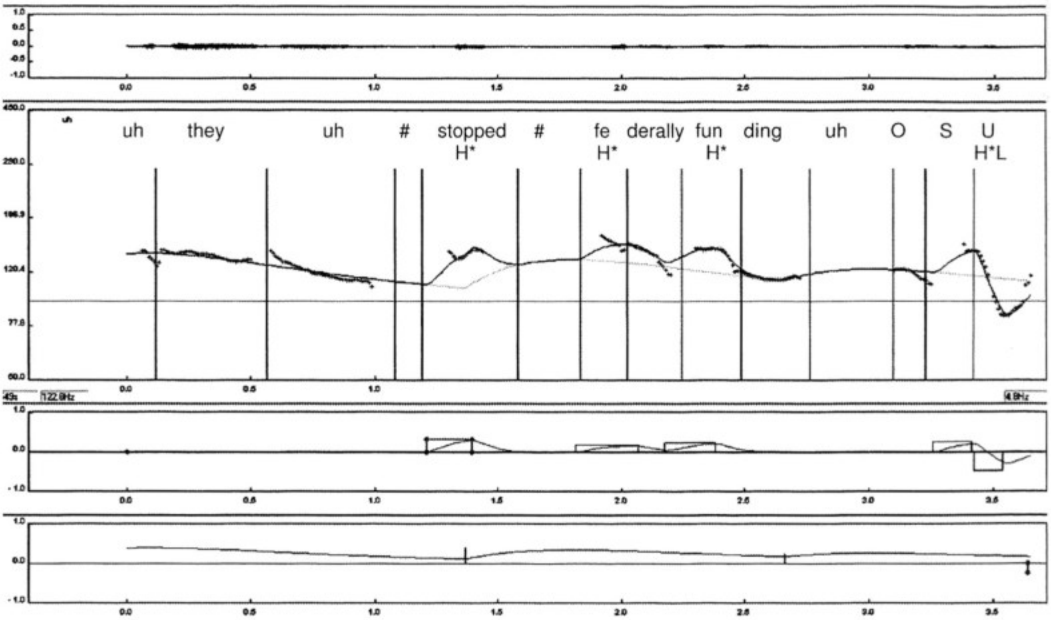

Figure 2 – Example of Fujisaki model-based analysis of the sentence "They stopped federally funding OSU". From the top to the bottom: The speech waveform, syllabic text and ToBI tone labels, the F0 contour (+ extracted and - modeled) and underlying accent and phrase components and commands.

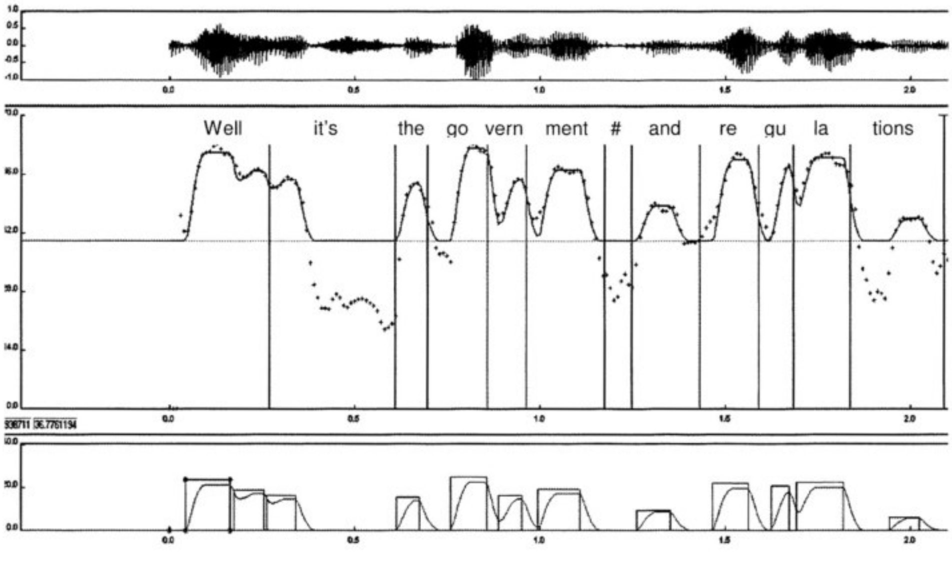

Figure 3 – Example of intensity modeling of the sentence "Well, it's the government and regulations..." From the top to the bottom: The speech waveform, syllabic text, the intensity contour (+ extracted and - modeled) and underlying syllable commands.

		Accent command			
		negative	none	positive	total
Accent label	none	26	1051	88	1165
	H*	0	57	330	387
	H*+L	5	0	21	26
	L*	12	36	36	84
	L*+H	0	0	43	43
	L+H*	0	0	29	29
	total	38	1144	547	1729

Table 1 - Correspondences between ToBI tone labels and accent commands (number of occurrences).

Accent label	$T1_{dist}$[ms]		$T2_{dist}$[ms]
H*	mean	-18	-47
	s.d.	88	133
H*+L	mean	-60	-214
	s.d.	48	123
L*	mean	-209	-250
	s.d.	131	105
L*+H	mean	185	82
	s.d.	88	115
L+H*	mean	10	-136
	s.d.	82	147

Table 2- Alignment of accent commands. T1 is related to the onset of the syllable ($T1_{dist}=T1-t_{on}$), T2 to the offset of the syllable ($T2_{dist}=T2-t_{off}$).

Table 2 displays the mean timing of positive accent commands depending on the tone label type. T1 is related to the onset of the syllable ($T1_{dist}=T1-t_{on}$), T2 to the offset of the syllable ($T2_{dist}=T2-t_{off}$). On the average, accent commands related to H* accents begin 18ms before the syllable onset and end 47ms before the syllable offset. As can be seen, the earliest alignment occurs in L* accents, followed by H*L and H*, whereas the accent command starts the latest in L*H accents. This corresponds to the expectation that a low syllables (L*, H*L), especially at the end of a phrase, is preceded by an accent command. In contrast L*H accents require a rise after the accented syllable nucleus and therefore a late accent command timing. Table 3 lists accent command amplitudes and syllabic durations depending on the type of the syllable. As can be seen, syllables preceding a boundary bearing and accent are the longest, followed by those with a boundary tone only and accent syllables in phrase-initial and medial positions. The mean accent command amplitudes are also higher for syllables bearing boundary tones than those that are accented in initial and final position. The corresponding figures for syllable command amplitudes are displayed in Table 4. On the average, accented syllables in phrase-medial and final position exhibited the highest intensities, followed by accented syllables phrase-finally.

Syllable type		Duration [ms]	Aa
Syllables with accent label, accent command and boundary tone	mean	364	.34
	s.d.	112	.17
	N	60	60
Syllables with boundary tone label	mean	297	.34
	s.d.	106	.22
	N	42	42
Syllables with accent labels and accent commands	mean	211	.32
	s.d.	85	.19
	N	399	399
Syllables without accent label but accent command	mean	212	.25
	s.d.	131	.18
	N	46	46
Unaccented Syllables	mean	162	-
	s.d.	93	-
	N	1165	1165

Table 3 - Accent command amplitudes and syllable durations depending on the type of syllable.

Syllable type	mean	s.d.	N
Syllables with accent label. accent command and boundary tone	18.70	4.79	60
Syllables with boundary tone label	16.16	4.03	42
Syllables with accent labels and accent commands	21.34	5.74	399
Syllables without accent label but accent command	16.94	7.95	46
Unaccented syllables	14.66	8.63	1165

Table 4 - Syllable command amplitudes depending on the type of syllable.

In order to determine whether the segmental onset of the syllable or the beginning of the syllable command was a better anchoring point for the accent command onset, we calculated $T1_{dist}$ with respect to both options. Histograms for $T1_{dist}$ calculated with respect to the segmental onset of the syllable (left) and with respect to the onset of the syllable command (right) are displayed in Figure 4. Both distributions are rather similar, just the mean is smaller for the syllable command onset as the reference (-60ms) than for the syllable onset (-8ms). This result is explained by the fact that the syllable command usually starts after the syllable onset. The standard deviations are rather similar (118ms vs. 121ms) for both cases, so the syllable command onset is not a more suitable anchoring point than the syllable onset.

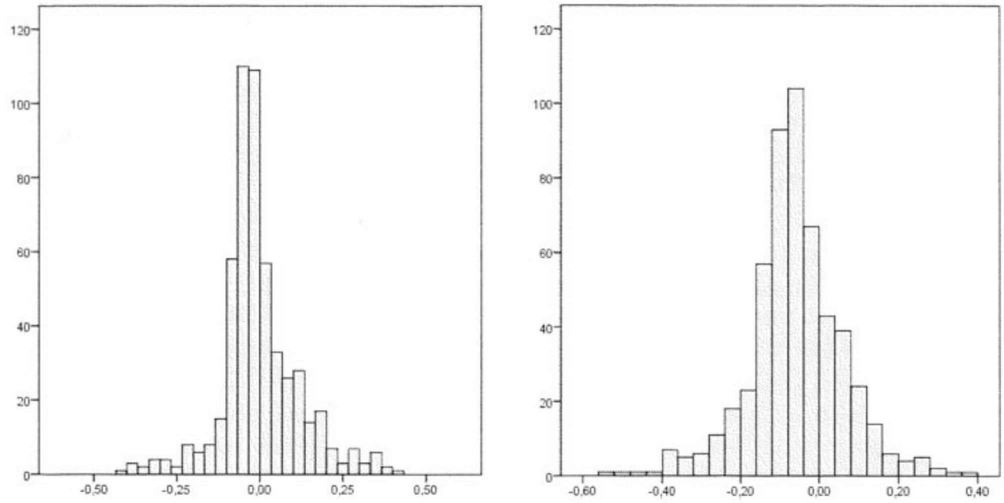

Figure 4: Histograms of $T1_{dist}$ calculated with respect to the segmental onset of the syllable (left) and with respect to the onset of the syllable command (right).

Finally, we examined the relationship between the duration of the syllable onset, nucleus and coda and the timing of the accent commands. To this end we correlated $T1_{dist}$ and $T2_{dist}$ with the duration of these syllable parts. The result is displayed in Table 5. The figures suggest that the longer the onset, nucleus and coda of a syllable are, the later the accent command starts. The situation is reversed for the relative accent command offset time: Whereas the duration of the onset does not have an influence, the longer the syllable, the earlier (with respect to the syllable offset!) does the accent command end. This indicates that the accent command once it is triggered continues for a certain amount of time that does not increase with the total duration of the syllable.

		duration onset	duration nucleus	duration coda
T1_dist	correlation	.237**	.218**	.119**
	significance	.000	.000	.006
	N	546	546	546
T2_dist	correlation	-.058	-.236**	-.293**
	significance	.173	.000	.000
	N	546	546	546

Table 5 – Correlations between accent command timing and the durations of onset, nucleus and coda of a syllable.

4 Discussion and Conclusions

This study examined the alignment of accent commands with the syllabic grid either represented by the segmental boundaries of syllables or by the onsets and offset of syllable commands which we calculated from intensity contours of a band-filtered version of the speech signal. As in earlier studies [6] on German we found that ToBI tone label classes exhibit specific timing characteristics of accent commands associated, for instance, early in H*L accents and late in L*H. Accent commands associated with phrase-final syllables and high boundary tones exhibit slightly higher amplitudes than phrase-initial or medial accent commands. The syllabic durations are also the highest in phrase-final position. Syllable command amplitudes are the highest in phrase-initial and medial accent syllables. The onset of the syllable command, however, does not seem to be a more reliable anchoring point for the accent command onset than the syllabic onset. The timing of the accent command onset is slightly delayed by long onsets or nuclei whereas the accent command offset basically follows the accent command onset after an almost fixed period of time and is not delayed by longer nuclei or coda.

It was observed that in the case of unvoiced consonants, especially in the syllable onset, the precise alignment of accent commands could not be established, due to a lack of F0 data points. Furthermore the intensity contour is just a coarse approximation – if at all – of the underlying articulatory gestures. Therefore future studies should employ motion capturing techniques that closely monitor tongue and jaw movements. All-voiced target utterances will facilitate an uninterrupted estimation of *F0* contours.

5 References

[1] 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.

[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] Mixdorff, H. and O. Jokisch, O.: Building an Integrated Prosodic Model of German. Proceedings of Eurospeech 2001, vol. 2, pp. 947-950, Aalborg, Denmark, 2001.

[5] Mixdorff, H., Fujisaki, H.: Production and perception of statement, question and non-terminal intonation in German. Proceedings of ICPhS, Stockholm,, Vol. 2, pp. 410-413, 1995.

[6] Mixdorff, H. and H. Fujisaki, H.: A quantitative description of German prosody offering symbolic labels as a by-product, Proc. of ICSLP 2000, vol. 2. Beijing, 2000.

[7] Pfitzinger, H.R., Mixdorff, H., Peters, B.: Correspondences between KIM-based symbolic prosodic labels and parameters of the Fujisaki model, Nordic Prosody X. pp. 261-272. Helsinki, 2009.

[8] Pierrehumbert, J.: The Phonology and Phonetics of English Intonation. Ph.D thesis, MIT, 1980.

[9] Pitt, Mark, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond: The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. Speech Communication, 45, 90-95, 2005.

[10] http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html

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

[12] Mixdorff, H.: A novel approach to the fully automatic extraction of Fujisaki model parameters. Proceedings of ICASSP 2000, vol. 3, 1281-1284, Istanbul Turkey, 2000.

[13] http://public.bht-berlin.de/~mixdorff/thesis/fujisaki.html.