THE INFLUENCE OF SYLLABLE STRUCTURE ON THE TIMING OF INTONATIONAL EVENTS IN GERMAN

Hansjörg Mixdorff and Hiroya Fujisaki †*

Dresden University of Technology*   Science University of Tokyo †

ABSTRACT

The present study deals with the influence of syllable structure on the the fine alignment of accent commands of the Fujisaki-model. The corpus used in this study consists of three-syllable words of German with word-accent on the second syllable which were uttered in citationform.

It is examined which factors influence accent command onset time $T1$ and accent command offset time $T2$. $T1$ can be predicted most accurately relative to the duration of the nuclear vowel. The prediction error can be further reduced when the type of the consonant immediately preceding the vowel is taken into account. $T2$ closely aligned with the segmental offset of the syllable.

1. INTRODUCTION

In earlier works it was shown that German intonation can be modelled as a series of rising and falling tone switches corresponding to the onsets and offsets of accent commands of the Fujisaki-model [1, 2, 3]. Perceptual comparison with other approaches has shown that this method produces high naturalness and a clear perception of intended accents [4].

In this approach accent commands were aligned with respect to the segmental onset of the nuclear vowel of the accented syllable.

This yields good results in certain syllable environments such as voiced consonant plus long vowel, but in cases of short vowels preceded by unvoiced consonant clusters, for instance, the accent is perceived as too weak and in some cases even sounds unnatural as it appears to be shifted towards the following syllable.

The current study aims at solving this kind of problem by examining more closely the fine temporal alignment of accent commands and accented syllables.

2. SPEECH MATERIAL

The speech material used in the study consists of utterances of three-syllable words (mostly verbs) by three native speakers of German, twice each. The words, altogether 67 tokens, exhibit the structure "be"-*accented syllable-plosive*-"en", for instance: "be-deu-ten" *"to signify"*, "be-fra-gen" *"to question"*, "be-stri-tten" *"denied"* etc. This kind of material was chosen in order to facilitate syllabic segmentation.

Table 1 gives an overview of syllable structures examined. Considering the great diversity of syllable structures in German this selection is far from complete.

Table 1: Syllable structures examined in the experiment.

syllable structure	number of tokens
V	1
C/V	8
C/C/V	10
C/C/C/V	1
V/C	2
C/V/C	31
C/C/V/C	9
V/C/C	1
C/V/C/C	3
C/C/V/C/C	1

In order to assess the microprosodic influence of the sounds involved on the F_0 contour, all tokens were uttered monotonously once.

3. METHOD OF ANALYSIS

The utterances were recorded on tape and converted at 16 kHz/16bit. The F_0 contours were extracted and analyzed using the Fujisaki-model by the method of Analysis-by-Synthesis. The speech segments of the accent syllables were labeled auditorily.

Figure 1 shows an example of analysis of the word "bemänteln"—*"to cover up"*. At the top the speech waveform is displayed. The curve drawn using + symbols indicates the measured F_0 contour, the solid line the F_0 contour produced by the Fujisaki-model and the dashed line its phrase component. The underlying accent command is displayed at the bottom. The vertical lines mark the boundaries of sound segments belonging to the word-accent syllable which have been SAMPA-labeled.

The F_0 contours of all tokens can be modelled using a single accent command and a single phrase command.

Figure 2 shows examples of accented and monotonous version of the word "bewegten" — *"moved"* by speaker HM. As can be seen from the monotonous version, the F_0 contour exhibits a downward trend at the [v]. In contrast, the F_0 contour on "en" does not show any influence by the preceding sound [t] in the monotonous version, hence the falling slope in the accented version is most likely caused by the falling slope of the accent command. This kind of effects must be taken into account when modelling the F_0 contours of the accented tokens, since the Fujisaki-model does not consider microprosodic influences.

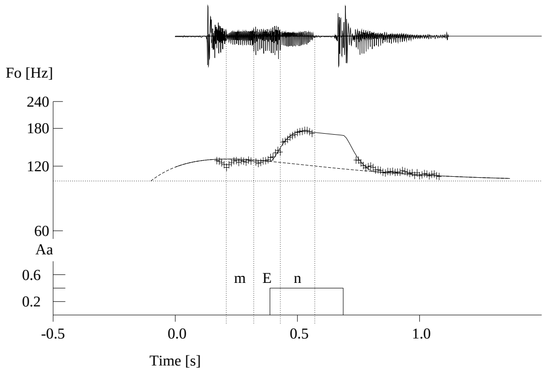

Figure 1: Example of analysis of the German word "bemänteln" — *"to cover up"*.

4. RESULTS OF ANALYSIS

Possible options for the alignment of accent command onsets and offsets with accented syllables included the alignment (1) relative to accent syllable durations, (2) relative to nuclear vowel durations and (3) absolute timing, i.e. the alignment with vowel or syllable onsets or offsets.

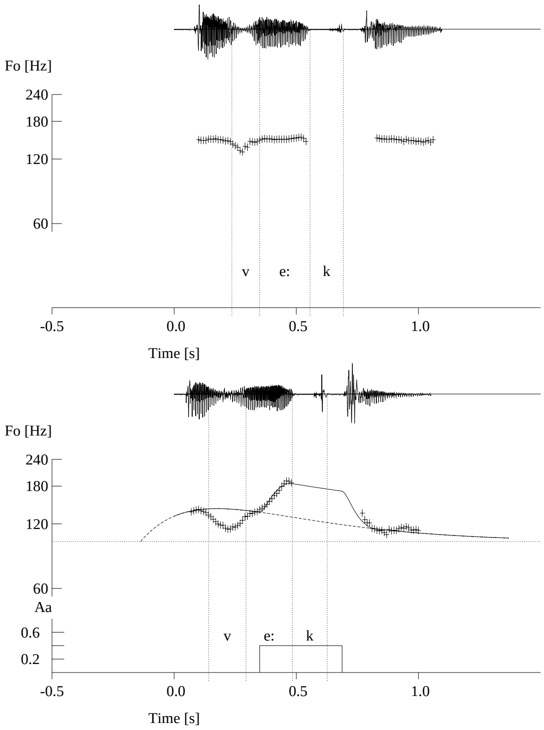

Figure 2: Examples of microprosidic effects on the F_0 contours: Dips caused by the sound [v]. Top: Monotonous version, bottom: Accented version

The duration of the consonant cluster preceding the nuclear vowel and the type of consonant immediately preceding the nuclear vowel were examined as further factors.

4.1 Accent Command Onset Time $T1$

In order to test the possible alignment options, the following parameters were determined:

$T1_{von}$ the delay between the onset time of the nuclear vowel and T1

$T1_{rsd}$ measured T1 expressed as a fraction of the accent syllable duration

$T1_{rvd}$ measured T1 expressed as a fraction of the nuclear vowel duration

t_{son} segmental onset time of the accented syllable

t_{von} onset time of the nuclear vowel

dur_s syllable duration

dur_v nuclear vowel duration

dur_c duration of consonant cluster preceding the nuclear vowel

Mean values and standard deviation for some of these parameters are displayed in Table 2.

Table 2: Mean square error for predicting $T1$.

Parameter	mean	std. deviation
$T1_{von}$	60 ms	60 ms
$T1_{rsd}$	51 %	11 %
$T1_{rvd}$	26 %	25 %
dur_s	420 ms	80 ms
dur_v	180 ms	80 ms
dur_c	160 ms	60 ms

Using these parameters, rules for predicting $T1$ were defined:

(1) relative to syllable durations: $T1_{p(1)} = t_{son} + mean(T1_{rsd}) * dur_s$

(2) relative to nuclear vowel durations: $T1_{p(2)} = t_{von} + mean(T1_{rvd}) * dur_v$

(3) absolute from vowel onsets: $T1_{p(3)} = t_{von} + mean(T1_{von})$

These rules were applied to the data in the corpus and the mean square error between measured and predicted $T1$ was calculated. Table 3 shows the results. It can be seen that the prediction relative to the duration of the nuclear vowel yields the best results. This means that $T1$ will be earlier in short vowels than in long ones.

Table 3: mean square error for predicting $T1$.

method	mean sq. error
rel. to vowel durations	2.31 ms^2
rel. to syllable durations	2.46 ms^2
abs. from vowel onsets	3.36 ms^2

Examining the dependency of $T1_{rvd}$ on the type of the preceding consonant, it is found that, though most consonant type-specific mean values for $T1_{rvd}$ are close to the total mean value, some differ considerably. These are listed in Table 4. The case [?] presents a nuclear vowel preceded by a glottal stop. When segmenting these cases the glottal closure was treated as part of the vowel which might explain the relatively late $T1$.

Table 4: Consonants with extreme $T1_{rvd}$.

Consonant	mean	std.deviation
[S]	-46 %	26 %
[R]	8 %	41 %
[n]	14 %	20 %
[m]	62 %	28 %
[?]	75 %	27 %

If $mean(T1_{rvd})$ is expressed as a function of the consonant type preceding the nuclear vowel, the mean square error for predicting $T1$ is further reduced to 2.12 ms^2. It was also found that dur_c had no significant influence on $T1$.

4.2 Accent Command Offset Time $T2$

The following parameters were extracted in order to determine the factors influencing $T2$:

$T2_{von}$ the delay between the offset time of the syllable and measured $T2$

$T2_{rsd}$ measured T2 expressed as a fraction of the accent syllable duration

$T2_{rvd}$ measured T2 expressed as a fraction of the nuclear vowel duration

t_{soff} segmental offset time of the accented syllable

In line with the considerations concerning $T1$, similar formulations for predicting $T2$ were derived and tested. Table 5 shows the results of the mean square error analysis indicating that $T2$ is closely aligned with the syllable offset. It must be stated, however, that in some cases, where the accent syllable-final consonants are voiceless, $T2$ cannot be determined exactly, unless inferred from the portion of the F_0 contour on the final syllable, as, for instance, in Figure 3.

5. PERCEPTUAL EVALUATION

For evaluating the effectiveness of the refined timing rules, an informal perception experiment

Table 5: mean square error for predicting $T2$.

method	mean sq. error
rel. to vowel durations	4.35 ms^2
rel. to syllable durations	2.28 ms^2
abs. from syllable offsets	2.24 ms^2

was conducted with synthesized stimuli produced with the TU Dresden TTS-system. Figure 3 shows examples of model-generated F_0 contours for the German words "bestatten"- *"to bury"* and "bemühten"- *"constrained"* which were produced following the refined alignment rules.

Four native speakers of German were offered pairs of stimuli, one stimulus produced with the original alignment rules and one with the refined ones, and had to decide which version they found more natural or if both were equally (un)natural. A slight preference for the latter stimuli was found, especially in the aforementioned problematic cases.

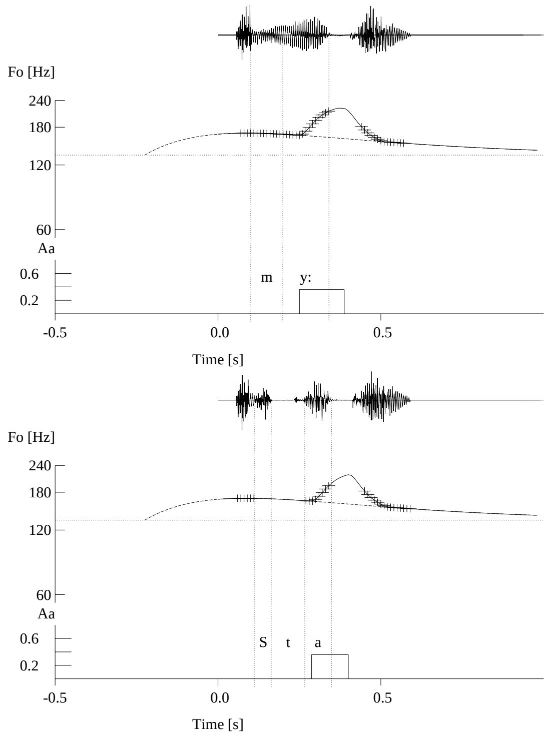

Figure 3: Examples of synthetic F_0 contours. Top: "bestatten" – *"to bury"*, bottom:"bemühten"– *"constrained"*.

Figure 3 shows examples of model-generated F_0 contours for the German words "bestatten"- *"to bury"* and "bemühten"- *"constrained"* which were produced following the refined alignment rules.

6. DISCUSSION AND CONCLUSION

The data used in this study is rather limited and only permits tentative conclusions. Rules where derived from and tested on the same material due to the relatively small amount of samples. Segmentation inaccuracies may have further blurred the results. Besides the speech material does not contain a complete selection of German speech sounds and sound combinations. Rules will have to be established for all types of possible leading consonants, as there is no evidence that they can be clustered into groups of similar sounds, such as nasals, unvoiced plosives, etc. The perceptual evaluation shows that considering the syllable structure when aligning accent commands with accented syllables yields better results than the original ad-hoc rules. Alignment may be more at variance in continuous speech than in utterances of citationforms. This will be a subject of further research.

REFERENCES

[1] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *Journal of the Acoustical Society of Japan (E), vol. 5, no. 4:* 233—241, 1984.

[2] Mixdorff, H. and Fujisaki, H. "Analysis of voice fundamental frequency contours of German utterances using a quantitative model", *Proceedings of the ICSLP '94, vol. 4*: 2231–2234, Yokohama, Japan, 1994.

[3] Mixdorff, H. and Fujisaki, H. " A Scheme for a Model-based Synthesis by Rule of F0 Contours of German Utterances", *Proceedings of the '95 Eurospeech, vol. 3*: 1823-1826, Madrid, Spain, 1995.

[4] Mixdorff, H. and Mehnert, D. "Perceptual Comparison of Three Different Approaches for Generating F0 contours in Text-to-Speech", *Fortschritte der Akustik, DAGA '98*, Zürich, Switzerland, 1998.