EFFECTS OF SYLLABLE STRUCTURE AND NUCLEAR PITCH ACCENTS ON PEAK ALIGNMENT: A CORPUS-BASED ANALYSIS

Bernd Möbius[1] *and Matthias Jilka*[2]

[1] Institute of Natural Language Processing; [2] Institute of English Linguistics
University of Stuttgart, Germany

moebius@ims.uni-stuttgart.de, jilka@ifla.uni-stuttgart.de

ABSTRACT

This paper describes the use of a unit selection corpus in carrying out an investigation of factors influencing specific aspects of the phonetic realization of tonal categories, concentrating on the alignment of peaks in H*L pitch accents in German. Three major linguistic parameters potentially influencing peak alignment are investigated. Two of them (syllable structure, nuclear pitch accents) are established influences while vowel quality is usually not considered relevant. Results from other studies are confirmed (peaks occur earlier in nuclear pitch accents, coda type influences peak position) and new findings are offered (in interaction onset type is more important than coda type). The presented procedure both describes the characteristics of the voice providing the corpus (allowing a more detailed phonetic realization of tonal categories, e.g., for speech synthesis) and offers general insights into which factors are relevant to the alignment of H*L peaks in German.

Keywords: intonation, peak alignment, German

1. INTRODUCTION

The present study continues an approach introduced in earlier research [4]. It uses a unit selection corpus, the IMS German Festival synthesis system [3], in order to conduct an investigation of factors influencing the phonetic realization of tonal categories. The synthesis system offers a voice database that includes a large number of individual instances of the phenomenon under investigation, in this case the alignment of F_0 peaks in H*L pitch accents in German.

A thorough description of the segmental and prosodic features of the linguistic units in which the H*L accents occur is part of the system as well. The approach should thus allow an effective determination of which phonetic realizations are most prototypical in certain well-defined contexts.

Ideally, this would not only lead to general insights into the phonetic realization of prosodic categories but, as the context information is available in a newly synthesized sentence as well, also to an improvement of prosody quality during the synthesis process - either directly via the unit selection itself or by means of subsequent prosodic modification.

The analysis presented here looks at two major linguistic factors potentially influencing peak alignment, namely syllable structure and the question of whether or not a pitch accent is nuclear. These factors have been the subject of a number of previous studies (e.g., [9], [2]) which, unlike this investigation, tightly controlled the segmental and prosodic environment of the peaks.

2. CORPUS

The speech database of the IMS German Festival synthesis system serves as a corpus for the investigation. It mainly consists of sentences that were selected from a newspaper corpus by means of a greedy algorithm in order to ensure good coverage. The corpus was recorded by a professional male speaker, contains approximately 160 minutes of speech (2601 utterances with 17489 words [7]) and was prosodically labeled using the GToBI System [6] (2681 instances of the H*L pitch accent).

3. PROCEDURE

The Festival synthesis system includes the "Festival feature functions" which can be used to describe a multitude of aspects of the segmental, syllabic, and prosodic structure of the utterances in the database [1]. The Festival framework is thus essential in defining the segmental and prosodic environment in which H*L peaks occur and allows itself to be extended with additional features that are deemed necessary.

The measurement of the peaks themselves is done automatically by locating the F_0 peak in a syllable labeled with a H*L pitch accent. For H*L the assumption that the peak is indeed in the same syllable is not problematic, but complications due to segmental effects cannot be avoided.

The general analysis of all labeled H*L accents must also disregard the fact that timing differences can either be phonetic or phonological in nature. As a consequence, differences in peak alignment that are not caused by the segmental and/or prosodic environment but are actually the expression of a different communicative function (as shown by [5] for early, medial, or late peaks in German) are not captured.

From the point of view of speech synthesis, this problem is not too pressing as the prediction of such differences in meaning is not yet possible anyway. Also, this kind of phonological variation is arguably less likely to occur in a corpus that mainly contains readings from newspaper articles.

It may be emphasized again that the approach taken in this and a previous study has the advantage of allowing for the effective analysis of large amounts of data. It does, however, not create a controlled environment in which influences from parameters other than the one investigated are excluded. This disadvantage can be dealt with by targeting interactions between specific parameters as demonstrated in the following analysis sections.

4. SYLLABLE STRUCTURE

Our investigation of the influence of onset and coda type on peak alignment was performed according to the classification established in [9], thus recognizing three main types: -V (voiceless obstruent), +V-S (voiced obstruent), +S (sonorant).

Earlier findings ([4]) confirmed that peak placement is significantly influenced by these factors. In the case of onsets, the peak is earliest when there is a sonorant in the onset (mean: 32.8% of syllable duration) and latest when a voiceless obstruent forms the onset (mean: 42.0%). If the onset consists of a voiced obstruent, peaks are generally located in-between (mean: 37.0%)

With different coda types, one can observe a significant difference of peak position between sonorant codas (mean: 41.7% of syllable duration) and codas solely made up of obstruents (mean 28.5% for voiced obstruents; 27.7% for voiceless obstruents).

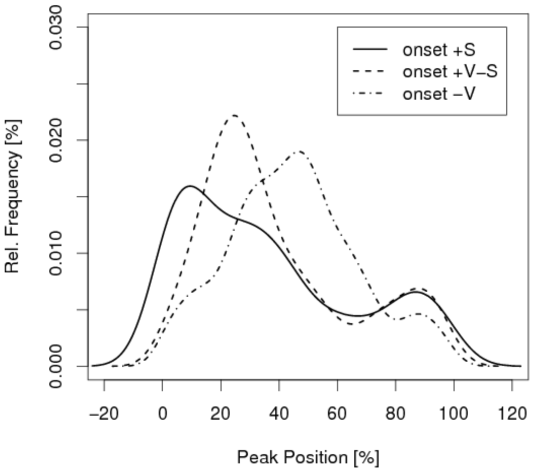

Figure 1: Density plot showing the relative frequency of peaks for the onset types sonorant (+S), voiced obstruent (+V -S) and voiceless obstruent (-V) with a sonorant coda.

4.1. Interaction of onset and coda types

The above investigations looked at the effect of a particular factor in isolation from other influences in the segmental and prosodic environment. For this reason, the analysis was extended to interactions between specific factors such as the combined influence of coda and onset.

If onset types are varied but coda type remains unchanged (sonorant), a significant movement of the peak can be observed (F [2, 2667] = 31.526, p < 0.001), as the peak occurs successively later when the onset is a sonorant (36.23% of syllable duration), a voiced obstruent (39.23%), or a voiceless obstruent (44.97%). The distinctive distribution of the three onset types is illustrated in the density plot in Fig. 1.

Variation of coda type (sonorant vs. voiceless obstruent) has a less distinct effect on the density distribution of the peak locations (see Fig. 2).

Most peaks apparently occur in the same locations. It must be pointed out, however, that the greater frequency of late peaks (manifested in the bump near the 100% syllable boundary) in sonorant codas, leads to a significantly later mean value (F [2, 2667] = 65.005, p < 0.001) for sonorant (36.23%) vs. obstruent codas (25.15%).

4.2. Two types of sonorant coda

In the Festival feature functions' classification of syllable structure types, coda type +S covers both closed syllables with actual sonorant coda consonants and open syllables. Differences between the two manifestations are, thus, to be expected.

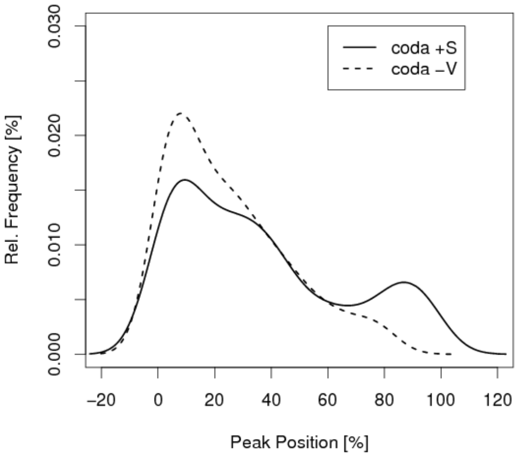

Figure 2: Density plot showing the relative frequency of peaks for the coda types sonorant (+S), and voiceless obstruent (-V) with a sonorant onset.

Examining the absolute interval between syllable start and peak location, it turns out that there is virtually no difference between open and closed (+S) syllables. For the former peak location is on average 97.59 ms after the beginning of the syllable compared to 97.24 ms for the latter.

As syllables with actual sonorant codas can be expected to be longer despite possible compensatory effects concerning vowel length (vowels in accented open syllables are unlikely to be short), this has the consequence that, in relation to syllable duration, peaks occur later in open syllables.

Indeed, in open syllables the mean value for peak location is 45.87% compared to 37.98 % for syllables with one sonorant coda consonant and 35.08% for syllables with two coda consonants. If the coda consists of only one obstruent, the peak occurs even earlier, at 30.83%. Fig. 3 attempts to visualize these results.

4.3. Interaction with vocalic features

The investigation of peak alignment depending on whether the accented vowel is tense (mean: 41.50%) or lax (mean: 37.34%) indicates the apparent significance ($p < 0.0001$) of that factor.

However, this effect is actually caused by the fact that, in the database, a much greater number of lax vowels (1132) is found in closed syllables than tense vowels (287). As indicated in section 4.2., peaks occur significantly earlier in closed syllables than in open ones.

Fig. 4 illustrates that the difference between tense and lax vowels in open (45.85% vs. 45.88%) and closed syllables (31.87% vs. 32.70%) by itself

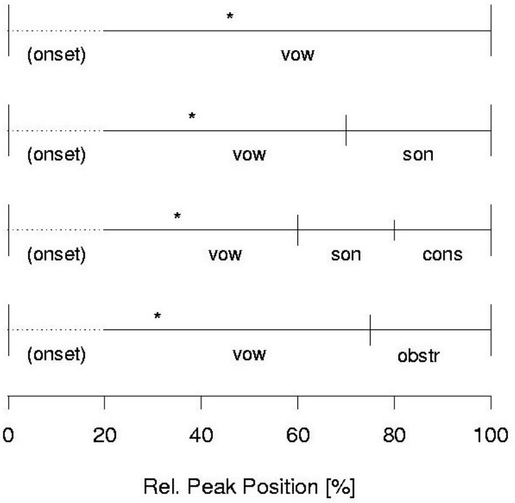

Figure 3: Relative peak position (*) for different coda compositions: open syllable (top row), coda with one sonorant consonant, coda with two consonants, coda with one obstruent consonant (bottom row)

is not significant and that the open/closed dichotomy is the truly distinctive factor.

Consequently, there is no reason to assume that tenseness is in any way relevant to peak position. Indeed, the example emphasizes that the methodology used in this study requires a certain awareness of the interactions between factors.

5. THE SPECIAL STATUS OF NUCLEAR PITCH ACCENTS

Various earlier studies, including work covering German (e.g., [2]), have demonstrated nuclear pitch

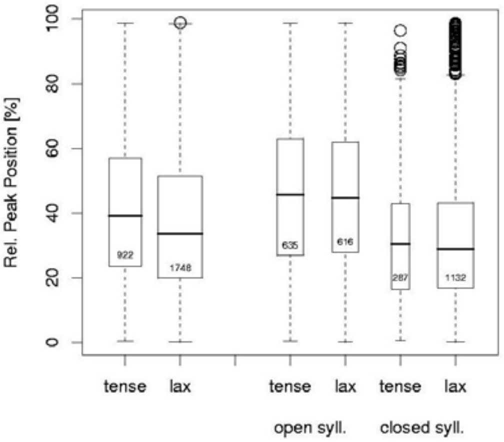

Figure 4: Comparison of peak alignment in tense vs. lax vowels. The apparent earlier peak placement in lax vowels (see left column) is due to the greater number of lax vowels in closed syllables. Figures within the boxes (and boxes' width) indicate number of instances in the corpus.

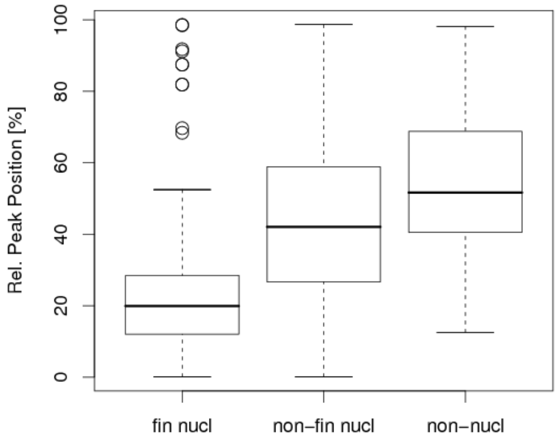

Figure 5: Boxplot showing relative peak position depending on type of pitch accent: nuclear PA on last syllable (median: 19.90%), nuclear PA not on last syllable (median: 42.05%), non-nuclear PA (median: 51.65%).

accents to be located earlier in the accented syllable than non-nuclear ones.

This result is confirmed also within the framework of the methodology and corpus presented in this study. Peak alignment in the final, i.e. nuclear, pitch accent of the intonation phrase is shown to occur significantly earlier (F [1, 2668] = 21.591, $p < 0.001$) than in non-final accents (mean: 38,48% vs. 53.43% of syllable duration).

There is an especially strong boundary effect of tonal repulsion (e.g. [8]) when the nuclear pitch accent is in the last syllable of the intonation phrase. In this case, the peaks are aligned rather early in the syllable (mean: 21.09%).

This brings up the question of whether nuclear pitch accents that are not on the last syllable and do not benefit from this effect are actually different from non-nuclear pitch accents. The data from the corpus indicates that they are, though to a slightly lesser degree ($p < 0.005$, $t = -3.1622$) with a mean peak location of 43.77% of the accented syllable. The results are summarized in Fig. 5.

The investigation thus shows that, even in a larger corpus with variable segmental and prosodic environment, it is possible to observe differences in peak alignment between nuclear and non-nuclear pitch accents.

6. CONCLUSION

Using the example of a narrowly defined problem, viz. examining selected influences of the segmental and prosodic environment on the alignment of German H*L peaks, this study aims to demonstrate the presented procedure's potential for a comprehensive examination of tonal events in a speech corpus. It is a large-scale approach that allows an effective investigation of a great number of instances of a particular phenomenon.

This is helpful for speech synthesis, as even very general measurements can be used as defaults to fall back on, should more complex rules not apply. In fact, for unit selection, the procedure offers the possibility of adapting to the potential prosodic idiosyncrasies of the individual speaker who provides the voice.

Despite an innate lack of control over the segmental and prosodic environment of an investigated peak, which prevents unambiguous identification of the effects of particular factors, the approach contributes important insights into various aspects of the phonetic realization of tonal categories. As concerns syllable structure, the significance of different coda types including the open/closed syllable dichotomy (see e.g., [9]) is corroborated along with the importance of different onset types (especially in interaction with coda types). The special status of nuclear pitch accents (peaks occur earlier than in non-nuclear pitch accents) is confirmed as well.

7. REFERENCES

[1] Black, A. W., Taylor, P., Caley, R., 1999. The Festival Speech Synthesis System – System documentation, CSTR Edinburgh. http://www.cstr.ed.ac.uk/projects/festival/manual/ - visited 26-Feb-07

[2] Grice, M., Baumann, S. 2002. Deutsche Intonation und GToBI. *Linguistische Berichte* 191, 267-298.

[3] IMS German Festival Homepage. http://www.ims.uni-stuttgart.de/phonetik/synthesis/index.html - visited 26-Feb-07

[4] Jilka, M., Möbius, B. 2006. Towards a Comprehensive Investigation of Factors Relevant to Peak Alignment Using a Unit Selection Corpus. *Proc. Interspeech* Pittsburgh, 2054-2057.

[5] Kohler, K. 1990. Macro and micro F_0 in the synthesis of intonation. In: Kingston, J., Beckman, M. (eds.), *Papers in Laboratory Phonology I.* Cambridge: CUP , 115-138.

[6] Mayer, J. 1995. Transcription of German Intonation – The Stuttgart System, Technical Report, University of Stuttgart

[7] Schweitzer, A., Braunschweiler, N., Dogil, G., Möbius, B. 2004. Assessing the Acceptability of the SmartKom Speech Synthesis Voice. *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 1-6.

[8] Silverman, K., Pierrehumbert, J. 1990. The timing of prenuclear high accents in English. In: Kingston, J., Beckman, M. (eds). *Papers in Laboratory Phonology I.* Cambridge: CUP, 72-106.

[9] Van Santen, J., Hirschberg, J. 1994. Segmental effects on timing and height of pitch contours. *Proc. ICSLP* Yokohama, 719-722.