Towards a Comprehensive Investigation of Factors Relevant to Peak Alignment Using a Unit Selection Corpus

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

[1] Institute of English Linguistics; [2] Institute of Natural Language Processing
University of Stuttgart, Germany
jilka@ifla.uni-stuttgart.de, moebius@ims.uni-stuttgart.de

10.21437/Interspeech.2006-403

Abstract

This paper aims to demonstrate the use of a unit selection corpus, the IMS German Festival synthesis system [1], in carrying out a comprehensive investigation of factors influencing specific aspects of the phonetic realization of tonal categories. The study restricts itself to the alignment of peaks in H*L pitch accents in German. First results confirm not only well-known effects of syllable structure, e.g., peaks occurring relatively early when there is a sonorant onset or relatively late when there is a sonorant in the coda, but also attest to the special status of the nuclear pitch accent vs. accents occurring earlier in the intonation phrase. Furthermore, instances of H*L in syllables directly at the phrase boundaries (initial or final) are shown to behave significantly differently from those that are located farther away. A similar effect is observed when another pitch accent follows the H*L peak in the very next syllable as opposed to a distance of two or more syllables. In these cases it also matters whether a low or high target is following (the peaks occur relatively later when followed by a L target). The results should have the benefit of both describing the specific characteristics of the voice providing the corpus (allowing a more detailed phonetic realization of tonal categories during the synthesis process) and offering general insights into which factors are relevant to the alignment of H*L peaks in German.

Index Terms: intonation synthesis, peak alignment, German

1. Introduction

The correct assignment and realization of tonal events is a persistent challenge to text- or concept-to-speech synthesis systems. When using a tone sequence-style approach like ToBI [2] for describing and modelling intonation, two major problem complexes are observed: predicting the overall intonation in terms of the choice of the appropriate tonal categories – which in many cases requires a certain "understanding" of the semantic, contextual circumstances - and the actual phonetic realization of such a tonal category once it has been determined.

The present paper addresses the latter problem and describes a first attempt at an analysis procedure that allows the examination of a large number of segmental and prosodic factors that could potentially be relevant to the specific phonetic manifestations of a particular tonal characteristic such as peak alignment. The information provided by the original voice database should allow an effective determination of which phonetic realization is most prototypical in a certain context. 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.

In this paper we hope to demonstrate an application of this idea with respect to a rather restricted area, that of peak location in the German H*L pitch accent. The data is provided by the corpus of the IMS German Festival synthesis system [1].

The investigation concentrates mainly on the influence of the pitch accent's position in the intonation phrase. Questions of interest are whether an instance of H*L is the first or last accent of the phrase as opposed to those instances where this is not the case, or whether it occurs in the first or last syllable of the intonation phrase or in another syllable. The influence of the distance and type of neighboring tonal targets (high or low) is also examined.

Additionally, we attempt to examine some of the factors regarding syllable structure and segmental structure previously described for American English (e.g., [3], [4]). This concerns the peak location in relation to different reference domains like the syllable or the rhyme and the importance of syllable composition (onset and coda types).

2. Corpus

As already indicated we used the speech database of the IMS German Festival synthesis system as a corpus for our investigation. It mainly consists of sentences that were selected from a newspaper corpus by a means of a greedy algorithm in order to ensure good diphone coverage.

As the synthesis system was part of the SmartKom multimodal dialog system [5] the database also includes material specific to the purposes of that project (place names, proper names, movies titles). The corpus was recorded by a professional male speaker, contains approximately 160 minutes of speech (2601 utterances with 17489 words [6]) and was prosodically labeled using the Stuttgart ToBI System [7] (2681 instances of the H*L pitch accent).

3. Procedure

The IMS German 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 [8]. Quite a number of the existing Festival functions (e.g., time of syllable start "syl_start", number of syllables to the next phrase break "syl_out") are very helpful in defining the segmental and prosodic environment in which H*L peaks occur. Additional features needed in order to look at circumstances expected to be promising were created as well (e.g., accent is in last syllable or not "syl_out_new", distance to the next intonation event is one syllable or more "dist_next_intevent_new" etc.). Statistical comparisons of relatively complex interactions of these features are carried out using R [9].

While the present analysis can thus account for a large part of the factors commonly assumed to affect alignment, the measurement of the peak itself is done in quite simple, automatic fashion by locating the F_0 peak in a syllable labeled with a H*L pitch accent. In the case of H*L the assumption that the peak is indeed in the same syllable is not problematic, but complications due to microprosody or voiceless regions 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. In the latter case this means that a difference in the alignment of the peak would not be the consequence of the segmental and/or prosodic environment, but actually the expression of a different communicative function (as shown by [10] for early, medial or late peaks in German).

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.

4. Analysis

4.1. Factors relating to syllable structure

The investigations presented in this section address issues that have been the subject of previous studies such as [3] and [4]. They deal with possible reference areas for the measurement of peak timing within the syllable on the one hand and the influence of the composition of onset and/or coda on the other hand. Contrary to those studies, there is, however, no control of the segmental and prosodic environment of the H*L peaks in the corpus (for example in [3] only H* peaks in the phrase-final syllable followed by a low phrase accent and a low boundary tone are considered) and the language examined language is German, not (American) English.

4.1.1. Reference points for measurement

One of the basic issues with respect to F_0 peaks is to look at their alignment in relation to different reference points in the segmental string. The Festival features allow the onset of the whole syllable, the onset of the vowel, the onset of voicing to act as such reference points. The peak position can be measured in absolute terms or relative to the respective reference area (syllable, rhyme, from start of voicing to end of syllable).

The results of such a measurement of the peak positions in all syllables labeled H*L in the database show that the majority of peaks is located early in the vowel (mean distance from syllable onset: 0.094 s (38.8% of syllable), from vowel onset: 0.011 s (6.8% of rhyme), from onset of voicing: 0.044 s (13.7% of area from start of voicing to end of syllable)).

This is also reflected in the density plot in Figure 1.

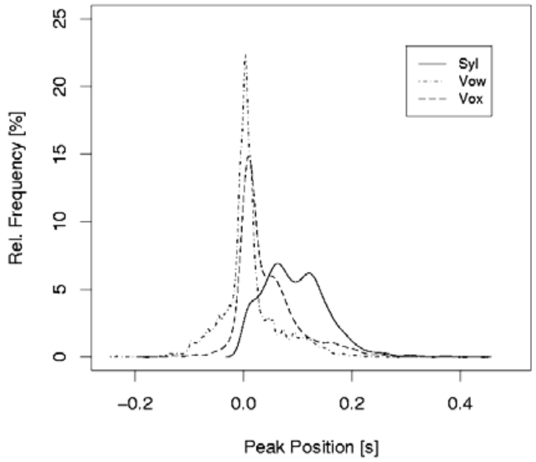

Figure 1. *Density plot showing the relative frequency of peak location in all 2681 instances of labeled H*L pitch accents. Peak location is indicated in terms distance (in s) from the onset of the syllable ("Syl"), the onset of the vowel ("Vow") and the onset of voicing ("Vox")*

The values for the different reference points do not, however, show convincing correlations. This may be explained by the syllable onset's importance to the start of the pitch movement as demonstrated in [4]. For this reason our presentation will concentrate on results in relation to syllable start/duration.

4.1.2. Effects of onset and coda class

The Festival features also allow a rapid investigation of the influence of onset and coda type on peak alignment according to the Van Santen/Hirschberg classification [3]: -V (voiceless obstruent), +V-S (voiced obstruent), +S (sonorant).

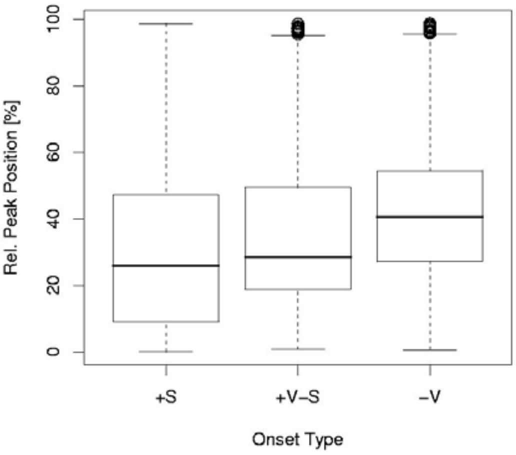

Figure 2. *Boxplot showing relative peak position depending on onset type: sonorant (+S, median: 25.95%), voiced obstruent (+V-S, median: 28.4%), voiceless obstruent (-V, median: 40.5%)*

Our findings confirm that peak placement (both with respect to syllable start and in relation to syllable duration) is significantly influenced by these factors.

As far as the three onset types (see also Figure 2.) are concerned, 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%).

For the different coda types (see Figure 3.) there is 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). The peak thus occurs clearly later when a sonorant coda is present, whereas there is no significant difference between the two obstruent classes.

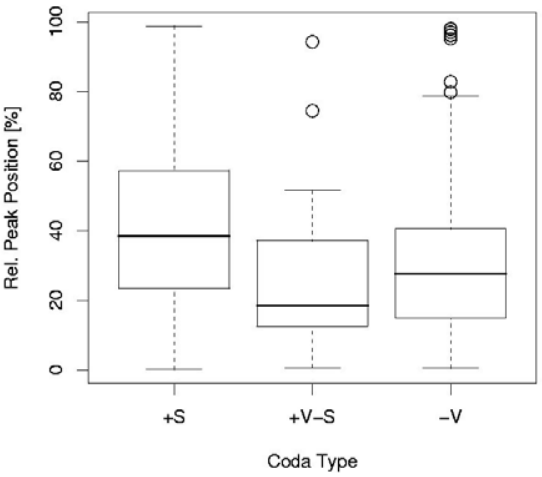

Figure 3. *Boxplot showing relative peak position depending on coda type: sonorant (+S, median: 38.4%), voiced obstruent (+V-S, median: 18.45%), voiceless obstruent (-V, median: 27.7%)*

4.2. Influence of the phrasal and tonal environment

The alignment of F_0 peaks is of course known to be affected by the proximity of other tonal events (pitch accents or phrase boundaries) which may lead to effects of tonal repulsion (e.g., [11], [12]). The addition of several new feature functions to the set of Festival features allows a comprehensive examination of a number of different constellations that occur in the corpus.

4.2.1. Position of the accented syllable in the phrase

A first interesting issue regarding the position of the H*L pitch accent in the intonation phrase is its distance to the preceding or following phrase boundary in number of accents. This amounts to the question whether it is significant if the pitch accent is the first, second, third, next-to-last, last etc. accent of the phrase. As determined by an analysis of variance this distance is shown to be significant (F [1, 2668] = 475.87, p < 0.001) when looking in the direction of the final phrase boundary. This is not the case with respect to the initial phrase boundary (F [1, 2668] = 0.2437, p = 0.6216).

Considering this result it is especially interesting to see whether pitch accents in the extreme positions of the phrase, i.e. the first or last (nuclear) accent, behave accordingly.

Indeed we find that peak alignment in the final pitch accent of the intonation occurs significantly earlier (F [1, 2668] = 21.591, p < 0.001) than in those accents which are not final (mean: 38,5% of syllable duration vs. 53.4% of syllable duration). Peak alignment in the first pitch accent of the phrase is on the other hand not significantly different (F [1, 2668] = 2.5009, p = 0.1139) from that of the other H*L pitch accents in the phrase.

The early alignment in nuclear pitch accents as opposed to non-final accents raises the question whether this is an effect that is facilitated by those instances that occur in the final syllable of the phrase and are thus being pushed forward by the following boundary tone. The comparison of H*L pitch accents in the phrase-final syllable with all other H*L pitch accents does in fact show a significant difference (F [1, 2668] = 496.56, p < 0.001). The peaks of final syllable accents are aligned quite early in the syllable (mean: 21.1% of syllable duration vs. 44.0%). The difference remains significant (F [1, 2668] = 104.98, p < 0.001) also when the nuclear phrase-final pitch accents are compared to nuclear pitch accents that are not in the phrase-final syllable (mean: 21.1% of syllable duration vs. 41.7%). In accordance with these results it is not surprising that similarly to a pitch accent's distance to the next phrase boundary in number of accents, its distance in number of syllables is also significant (F [1, 2668] = 420.01, p < 0.001), and that, correspondingly, there are no significant results with regard to distance to the preceding phrase boundary (F [1, 2668] = 0.0194, p = 0.8892).

4.2.2. Influence of neighboring tonal targets

If boundaries can affect peak alignment, then it is of course reasonable to assume that other neighboring tonal events, i.e., pitch accents, will do so as well. In a comprehensive analysis of the influence of such adjacent tonal events three major questions are of interest, namely whether the adjacent target is high or low, whether it is preceding or following and how far in terms of number of syllables it is away from the examined pitch accent.

High (H) or low (L) targets are defined by the target point closest to the examined pitch accent, a preceding L*H pitch accent would thus be registered as H, a following one as L.

A first analysis that disregards the distance of a neighboring target to the accented syllable shows that peak alignment is not influenced by the type of target preceding it (F [1, 2461] = 1.643, p = 0.2000). It is, however, of weak significance whether a H or L target follows (F [1, 2665] = 6.0593, p < 0.05; mean peak position when H target follows: 36.4% of syllable duration vs. mean peak position when L target follows: 39.3% of syllable duration).

Even if distance - measured as a comparison of tonal events which are one syllable away from the examined pitch accent with all others that are farther away - is added as a factor, type of preceding target is not relevant to peak alignment (F [1, 2461] = 1.5438, p = 0.2142). For type of following target on the other hand it is (F [1, 2665] = 80.584, p = 0.001), as alignment occurs significantly earlier when the next tonal target follows immediately in the next syllable (mean peak position: 33.6% of syllable duration vs. 42.2% of syllable duration).

This result is also confirmed when comparing the influence of H and L targets at a distance of either one or two syllables from the accented syllable. Here, peak alignment is significantly different for all four possibilities (mean H following after 1 syllable: 26.1 %; mean H following after 2 syllables: 39.3 %; mean L following after 1 syllable: 34.5% mean L following after 2 syllables: 43.2%). In this case there is thus also a difference depending on whether a high or low target is following (see also Figure 4.)

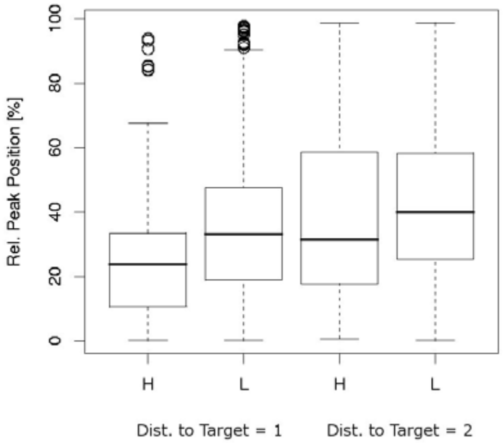

Figure 4. *Boxplot showing relative peak position depending on distance and type of following target:H target after 1 syllable (median: 23.8%), L target after 1 syllable (median: 33.2%), H target after 2 syllables (median: 31.4%), L target after 2 syllables median: 39.9%),*

5. Conclusions

While it deals with a relatively narrowly formulated problem (examining selected influences of syllabic and tonal environment on peak alignment in German H*L pitch accents), this paper gives an impression of the presented procedure's potential for an effective and comprehensive examination of virtually all aspects of the tonal events present in a speech corpus. The procedure can be extended to include more contextual factors (in terms of additional Festival functions) and apply to other points in the contour that are deemed to be important, e.g. valleys, anchor points. Correspondingly, other phonetic manifestations of various tonal categories such as peak height or contour shape can be examined as well.

The presented analysis must therefore remain somewhat superficial: the effects of individual factors are studied separately, their interactions (e.g. peak alignment in a phrase-final syllable with a sonorant coda) in a large speech corpus are not considered to a great extent, while on the other hand it is also not possible to control the segmental and prosodic environment in such a way that the effect of one particular parameter can be identified unambiguously because other influences are excluded.

A large-scale approach that looks at a great number of instances of a certain phenomenon, in this case peak location in H*L pitch accents, can nevertheless offer interesting insights into which contexts/environments are of significance. The present study did confirm syllable-dependent factors such as the relevance of onset and coda type as well as prosodic factors like the special status of nuclear pitch accents. It was also found that peak alignment is regularly pushed forward by immediately following phrase boundaries or other tonal events. Interestingly, the effect was stronger when the following tonal target was high.

With respect to speech synthesis it can be added that even very general measurements such as the most frequent peak location in all H*L pitch accents of the corpus may have their use 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.

6. References

[1] "IMS German Festival Homepage," [http://www.ims.uni-stuttgart.de/phonetik/synthesis/index.html]

[2] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J., "ToBI: A standard for labelling English prosody", *ICSLP Proc.*, 867-870, 1992

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

[4] Van Santen, J. and Möbius, B. "A Quantitative Model of F0 Generation and Alignment". In Botinis, A. (ed.), Intonation – Analysis, Modelling and Technology, Kluwer, Dordrecht, 269-288, 2000

[5] Wahlster, W. (ed.), SmartKom: Foundations of Multimodal Dialogue Systems, Springer, Berlin, 2005

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

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

[8] Black, A. W., Taylor, P. and Caley, R., "The Festival Speech Synthesis System – System documentation", CSTR Edinburgh, 1999 [http://www.cstr.ed.ac.uk/projects/festival/manual/]

[9] "The R Project for Statistical Computing" [http://www.r-project.org]

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

[11] D'Imperio, M. "Italian Intonation: An Overview and some Questions" *Probus 14(1)*, 37-69, 2002

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