Quantal effects in the Temporal Alignment of Prosodic Events

Grzegorz Dogil and Antje Schweitzer

Institute for Natural Language Processing, Stuttgart University, Germany
{dogil,schweitzer}@ims.uni-stuttgart.de

ABSTRACT

We present a method for investigating the temporal alignment of intonation events by parametrizing F0 contours. Results for three German single-speaker corpora and one American English multi-speaker corpus show that the speakers generally avoid to place peaks in syllable onsets. We suggest that this is a quantal effect [9] which results from the fact that syllable onsets are boundaries in tonal production.

Keywords: Prosody production, tonal alignment, F0 approximation

1. INTRODUCTION

Quantal Theory [9] claims that nonlinearities in the articulator-to-acoustic mapping are responsible for quantal effects that are the basis of distinctive features in speech perception. Consider the following example, which is discussed in detail in [4]: Small movements of the articulators usually correspond to small changes in formant structure. But if the movement causes the second formant to lie in the area of the second subglottal resonance, the change in formant structure is more severe: The pole caused by the subglottal resonance introduces discontinuities in that formant, which may degrade perception. [9, 4] argue that the subglottal resonance constitutes a *boundary* which is generally avoided in production. The boundary causes a quantal effect: In production there are configurations in which the second formant lies below the second subglottal resonance, and configurations in which it lies above this resonance, thus the boundary effectively splits the vowel formant space in two. [9, 4] claim that the boundary serves as an acoustic implementation of the distinctive feature $[\pm back]$ in vowels.

Further instances of boundaries according to [9] are caused by nonlinearities such as (i) the size of the constriction in consonant production and its relationship to the amplitude of the resulting turbulence noise, (ii) the size of the velopharyngeal opening in the production of nasal consonants and its relationship to the influence of the spectral characteristics of the oral cavity on the resulting speech signal, and (iii) the tong blade position in obstruent consonants and its relationship to the frequency of the spectrum prominence.

We suggest that such boundaries exist in the prosodic domain as well. We will introduce a methodology for investigating peak alignment in intonation contours (sec. 2) and present results for three German corpora and one American English corpus (sec. 3). Finally, it will be discussed that the results can be interpreted as evidence for quantal boundaries in prosody production (sec. 4).

2. METHOD

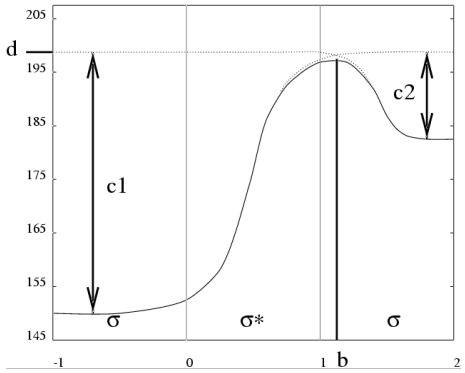
We analyze the temporal alignment of prosodic events in several corpora by way of parametrizing F0 contours syllable by syllable using the PaIntE method described below. Before approximation, the contours are smoothed using a median smoother and interpolated across unvoiced regions. Smoothing is intended to eliminate most microprosodic effects.

The parametrization yields 6 parameters for each syllable, one of which determines the temporal alignment of an F0 peak with the syllable structure. We analyze the distributions of this parameter in several corpora. For two of these corpora, pitch accent type labels are available, which makes it possible to consider distributions of the alignment parameter by pitch accent type. For one corpus, only pitch accent location is available; thus we can investigate the distributions for pitch accented syllables in general, not differentiating between accent types. For one corpus, we have no prosodic labels at all, thus we investigate the parameter distribution across all syllables.

2.1. PaIntE Approximation

The PaIntE ("Parametrized Intonation Events") model [7, 6] parametrizes intonation contours using six linguistically motivated parameters. A diagram of the PaIntE approximation function is given in fig. 1. It approximates the F0 contour related to a syllable in a window including the syllable and its two neighbors, as indicated by the three intervals separated by vertical lines in fig. 1. The time axis is normalized in a way that the syllable boundaries occur at integer values, with the syllable of interest stretching from 0 to 1, as illustrated in fig. 1.

Figure 1: PaIntE approximation function reproduced from [7]. The approximation window represents three syllables. The syllable of interest is indicated by the asterisk (σ*). See text for further details.



The exact contour is determined by 6 parameters: Parameters *a1* and *a2* (not depicted in fig. 1) represent the steepness of rise and fall, respectively, while *c1* and *c2* specify the corresponding amplitudes. Parameter *d* can be interpreted as approximating the absolute peak height in Hertz, and *b* determines the temporal alignment of the peak. The remainder of this paper will focus on this last parameter.

For the analyses presented here, we used a modified version of the *anchornorm* normalization described by [6], who normalized the time axis with regard to syllable structure by splitting each syllable into three parts representing the unvoiced onset, the voiced onset and nucleus, and the coda. The unvoiced onset was adjusted linearly to 0.3 times the syllable duration, the voiced onset and nucleus to 0.5, and the coda to 0.2. We modified this procedure to map the onset, regardless if voiced or unvoiced, to 0.3, the nucleus to 0.5, and the coda to 0.2. Thus, *b* values between 0 and 0.3 indicate that the peak occurred within the onset of the center syllable; values of 0.3 to 0.8 indicate its location in the nucleus, and values between 0.8 and 1.0 indicate that the peak was in the coda; analogously, values between -0.7 and -0.2 indicate the peak occurred in the nucleus of the first syllable, etc.

2.2. Data

2.2.1. German unit selection corpora

We used a male database (28,000 syllables, 14,000 words, and 6,000 pitch accents) and a female database (34,000 syllables, 17,000 words, and 8,000 pitch accents). Both were originally used for unit selection speech synthesis. The utterances represent typical utterances of 5 different genres. They were manually prosodically labeled according to GToBI(S) [5].

2.2.2. American English Switchboard corpus

Here we used the Switchboard corpus of telephone conversations between non-professional speakers, which is annotated for accent location [1]. Two dialogs had to be excluded because the approximation failed at some point in the dialogs. This left 73 dialogs by more than 100 speakers, adding up to more than 8 hrs. of dialog, with approx. 122,000 syllables, 96,000 words, and 33,000 pitch accents.

2.2.3. German audio book corpus

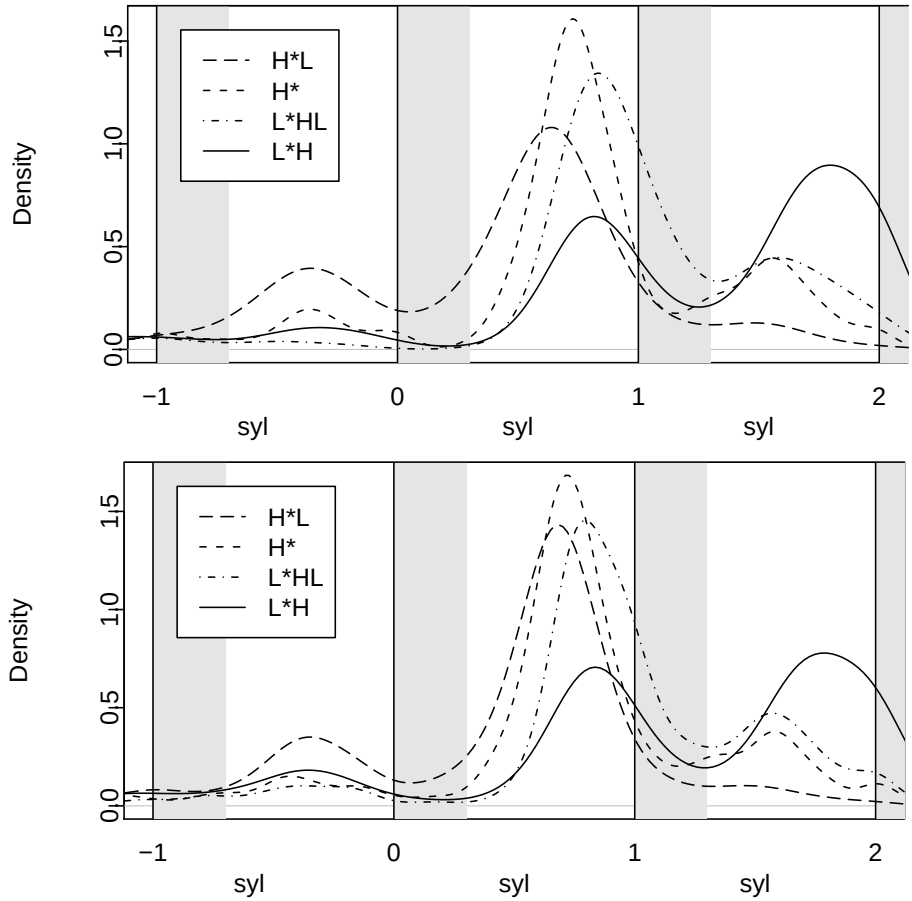
The fourth corpus consists of audio recordings from the LibriVox open-source audio book project [3], in which volunteers provide recordings of freely available books. We chose recordings of Gottfried Keller's "Kleider machen Leute" (*Fine feathers make fine birds*) and of seven chapters of E.T.A. Hoffmann's "Elixiere des Teufels" (*The Devil's Elixir*), all read by the same non-professional speaker, adding up to approx. 4.5 hours. Annotation was done automatically by forced alignment on the segment, syllable, and word level yielding roughly 75,000 syllables and 46,000 words. There was no prosodic annotation.

3. RESULTS

We used R's [8] kernel density estimates to calculate the likelihood of the observed values of pitch accent alignment as depicted by parameter *b* for each of the corpora described in sec. 2. The results for the unit selection corpora are presented in fig. 2. Again, the x-axis is the temporal axis, normalized for syllable structure. Syllable boundaries are marked by vertical lines, syllable onsets by gray bars. Despite the similarity to fig. 1, the lines in fig. 2 do not correspond to F0 contours; they are estimated densities which indicate the likeliness of the *b* values. Thus, peaks in fig. 2 do not correspond to F0 peaks; instead they occur at the point where F0 peaks were most likely.

Results for the male speaker are in the upper panel, results for the female speaker in the lower one. The densities were calculated separately for each of the most frequent accent types in the corpora. They indicate that both speakers realize the peaks in H*L and H* accents (long-dashed and short-dashed lines) approx. in the middle of the rhyme of the accented syllable (the white part of the center syllable in fig. 2). In L*HL accents (dot-dashed lines), both speakers tend to realize the peaks slightly later, toward the coda. Even though the likelihood of the peak decreases in the course of the onset of the following syllable (the dot-dashed line in fig. 2 falls throughout the gray onset part of the last

Figure 2: Kernel density plots for the *b* parameter for the most frequent pitch accent types in the male (upper panel) and female (lower panel) unit selection corpora. F0 peaks are generally avoided in syllable onsets (gray bars).

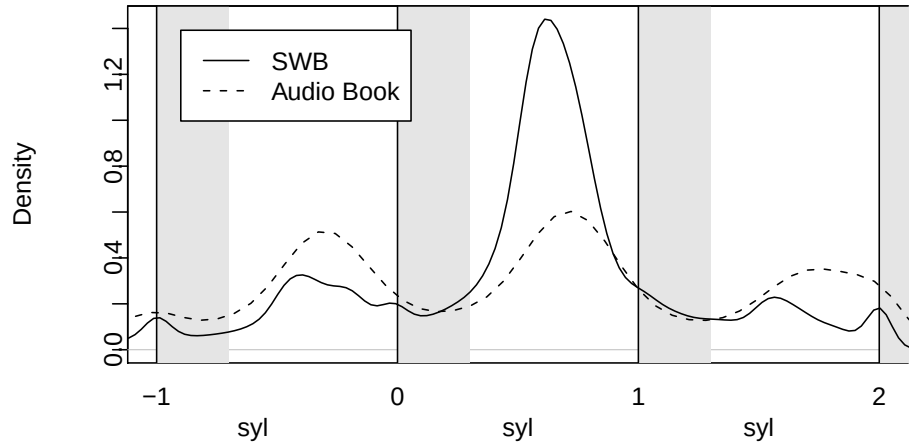


syllable), peaks in L*HL accents may occasionally occur even in the onset. In L*H accents (solid lines), the peak is either on the center syllable or on the following syllable, in both cases roughly at the end of the nucleus part. It is evident when looking at the results in fig. 2 that peaks in general very seldom occur in syllable onsets in these data: for all accent types and for both speakers, the distributions exhibit pronounced valleys at the gray areas which correspond to syllables onsets, even for L*HL accents.

The results for the American English Switchboard corpus (fig. 3) are similar (solid line, "SWB"). As the corpus only provides labels for accent location but not for accent type, we can just investigate the density of accented syllables in general. They are most likely to have their F0 peak in the middle of the rhyme of accented syllable. It is interesting to note that, in contrast to German, for accented syllables peaks on neighboring syllables are very unlikely in the Switchboard data. For instance, in the German data, L*H accents were most likely to have their peak in the rightmost syllable, and this was also the case for a few H* and L*HL accents. In the Switchboard data, F0 peaks on the rightmost syllable are the exception, as evidenced by the very small maxima at these points in the distributions. We would expect that such exceptions are related to L*+H, L*+!H, and possibly L*, accents. Since these are indeed rather infrequent in American English, this could account for the fact that peaks in the center syllable dominate in the Switchboard data.

The results for the German audio book corpus are indicated by the dashed line in fig. 3. Since prosodic labels were not available for this corpus, there is only one distribution representing all syllables in the corpus, irrespective of their prosodic properties. F0 peaks were similarly likely on any of the three syllables. This is because the parametrization has been carried out for all syllables, including unaccented syllables with possibly neighboring accented syllables, and in these cases, the F0 peak related to the neighboring accented syllable is captured in the pa-

Figure 3: Kernel density plots for the *b* parameter in Switchboard accented syllables (solid line), and for all syllables in the audiobook corpus (dashed line). F0 peaks are generally avoided in syllable onsets (gray bars).



rameters of the unaccented middle syllable. Apart from this aspect, the distribution again exhibits valleys at the syllable onsets, confirming that F0 peaks in onsets were very unlikely.

4. DISCUSSION

All data presented here confirm that speakers both in the American English spontaneous conversations and in the German read speech corpora consistently avoided placing F0 peaks in syllable onsets. This is reminiscent of House's [2] model of optimal tonal perception, which claims that tonal movements through areas of maximum new spectral information and intensity change (such as syllable onsets) are perceived as level tones rather than as tonal movements. The model states that gestures of tonal movement must be synchronized to occur after these areas in order to be perceived as such.

However, our data seem to suggest that it is not necessarily the movement which is timed to occur after the onset; rather, it is the peak which occurs usually at the same point in the syllable, viz. in the middle of the rhyme. For the two German corpora for which we have accent type labels available (cf. fig. 2), we can observe that the peak is approx. at the same point in falling (H*L) accents as in high (H*) accents (the short-dashed and the long-dashed lines have their maxima roughly at the same point in the course of the center syllable). If the fall in H*L was timed to start after the onset, we would expect the peak right at the beginning of the nucleus. Admittedly, rising (L*H) accents and rise-falls (L*HL) have their peaks a little later, after the middle of the rhyme; however, if speakers were to time the rise so it is realized to start after the onset only, we would expect the peak to be reached even later. However, the PaIntE model does not provide information about the start of the rise, so this question can not be answered exhaustively here.

To summarize, we consistently observe the same tendency for all corpora: peaks are realized either before or after syllable onsets. This quantal effect—"either before or after"—supports our hypothesis that syllable onsets are boundaries in prosody production, similar to the way subglottal resonances are boundaries in vowel production.

5. REFERENCES

[1] Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D. 2010. The NXT-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

[2] House, D. 1996. Differential perception of tonal contours through the syllable. In *Proc. ICSLP* Philadelphia, vol. 1, 2048–2051.

[3] LibriVox—acoustical liberation of books in the public domain. http://www.librivox.org

[4] Lulich, S. M. 2010. Subglottal resonances and distinctive features. *Journal of Phonetics*, 38(1):20–32.

[5] Mayer, J. 1995. Transcription of German intonation—the Stuttgart system. Technical report, IMS, University of Stuttgart.

[6] Möhler, G. 2001. Improvements of the PaIntE model for F0 parametrization. Manuscript.

[7] Möhler, G., Conkie, A. 1998. Parametric modeling of intonation using vector quantization. In *Proc. 3rd Int'l Workshop on Speech Synthesis* Jenolan Caves, 311–316.

[8] R Development Core Team 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[9] Stevens, K. N. 2003. Acoustic and perceptual evidence for universal phonological features. In *Proc. ICPhS* Barcelona, 33–38.