Acoustic correlates of word stress in German spontaneous speech

Grigorij Aronov, Antje Schweitzer

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

grigorij.aronov@ims.uni-stuttgart.de, antje.schweitzer@ims.uni-stuttgart.de

Abstract

The acoustic properties of word stress have been explored in a number of studies. However, there is little research on German word stress, and even less on its realization in spontaneous speech. This paper tests whether parameters that have been found to implement word stress in mostly laboratory speech are also employed in a corpus of German spontaneous speech. Specifically, we consider spectral tilt, syllable duration and pitch. While the results for syllable duration conform with the prevalent finding that stressed syllables have a higher duration, we find no significant effect of pitch. In the case of spectral tilt however, we observe contradicting results, depending on the way we quantify tilt.

Index Terms: word stress, spectral tilt, spectral balance, intensity, pitch, spontaneous speech, syllable duration

1. Introduction

In this paper we explore the effects of word stress on spectral tilt, syllable duration and pitch. Word stress (also called *lexical stress*) denotes a relation of prominence between emphasized and unemphasized syllables of a word. In fixed-stress languages, there are constraints regarding the position of stressed syllables in words; for example, Turkish is considered a language with word-final stress, and Hungarian one with initial stress. Some languages are claimed to have no word stress at all, for instance French [1] or Chinese [2]. German on the other hand, just like English, has variable word stress, the position of which has to be learned together with the pronunciation of a word. Thus, identifying the stressed syllable can aid in word recognition. Speakers are expected to mark stress in production acoustically, and listeners rely on these acoustic cues to detect stressed syllables.

Since the 1950's, a considerable body of research on word stress has identified parameters that are employed in speech perception and speech production to detect or to mark word stress, however with inconsistent results. [3, 4, 5] found that duration, F_0, vowel quality and intensity affect the perception of stress in English listeners. Duration proved to be a stronger cue than intensity. [6] for Dutch also investigated duration, vowel quality, and intensity; however, they calculated the intensity both as overall intensity over the whole spectrum and as the individual intensities in several frequency bands. They confirmed that duration is a strong correlate in the production of stress. In addition they found that increased overall intensity is a poor cue, while increased intensity in the higher frequency bands is a reliable cue. This established that stressed syllables are not characterized by overall greater amplitudes, but that there is a shift in what [6] call the "spectral balance" of stressed syllables, and they explain this shift by greater vocal effort. A follow-up perception study [7] confirmed the perceptual relevance of these parameters.

However, using a rather small English corpus [8] failed to reproduce the results of [6], who had used a Dutch corpus. While [8] provide further evidence for spectral balance differences in vowels produced with and without pitch accents, no difference could be found between stressed and unstressed syllables when pitch accent was not involved. Furthermore, measurements of duration were inconsistent.

[9] compared English dialects regarding correlates of word stress. They built a classifier to predict human judgments of stress in a large corpus of natural speech, comparing acoustic correlates by their predictive power. These correlates were loudness, aperiodicity, spectral slope, several features related to F_0, and a running measure of duration (measuring how long acoustic properties remain stable). While in the literature F_0 is often considered to be a strong cue, [9] provide evidence that F_0 is a weak cue for prominence as neither local F_0 changes, values nor variances were particularly predictive. Also, their results did not support the importance of spectral tilt. Instead, loudness and duration turned out to be the primary indicators with loudness being more important. However, [9]'s measure of spectral slope was different from the way spectral balance was quantified by [6].

In fact, in the literature there are several different approaches to measure relative intensities in the spectrum. Further alternatives include calculating the difference between the amplitude of the first harmonic and the third formant [10], or the difference between the first and the second harmonic [8]. In this paper we tested an alternative where we quantified the spectral balance as the slope of a linear regression line of a spectrum using Praat [11]. This is similar to the measure used by [9]. We will use the term *spectral tilt* to distinguish methods that make use of the regression line fitted to a spectrum from other methods that look at differences between specific regions or points of interest, such as frequency bands, or specific harmonics or formants. We will refer to the latter by the term *spectral balance*.

The aim of this paper is to investigate the implementation of word stress in German on a corpus of natural speech. Most earlier studies are concerned with the analysis of simple, separate, sometimes novel words [10] or distinct sentences. This facilitates research greatly as noise is reduced that way. However, the results mentioned above were not tested on real everyday speech. Moreover, little research can be found on word stress in German. In this paper we will use a large corpus of spontaneous speech to address both problems. We will analyze distributions of syllable duration, vowel pitch, spectral tilt and spectral balance for stressed vs. unstressed syllables and compare our results to the results found for other languages in the literature.

2. Data and feature extraction

GECO (**GE**rman **CO**nversations) [12] is a database consisting of 46 fully spontaneous dialogs, each with a duration of approximately 25 minutes resulting in 20.7 hours of dialog (two channels), with $\sim$ 250,000 words, making it the largest German database of its kind to the best of our knowledge. It was annotated on the segment, syllable, and word levels by forced alignment. Word stress is annotated as part of the syllable annotation in the forced alignment process, and whether a syllable is stressed or not is determined by way of a lexicon look-up.

For each of the approx. 310,000 syllables in the corpus, we extracted its stress and its syllable duration from the annotations, as well as pitch values at mid vowel using get_f0 from the ESPS software package. For every vowel for which we had more than 5 voiced frames (approx. 41 % out of 310,000), we used Praat to calculate spectral tilt between 0 to 5,000 Hz using the “Report spectral tilt” function on a long-term average spectrum with frequency bins with bandwidths of 100 Hz. Since the values for tilt obtained in this way are usually negative (due to the overall falling tendency of the spectrum), we multiplied all values by -1 to obtain positive values if the spectrum is falling. Thus higher values represent higher (negative) tilt. We also calculated spectral balance for each vowel by taking the absolute difference between the mean intensity in two frequency bands B1 (0-0.5 kHz) and B2 (0.5-1.0 kHz), in analogy to the spectral balance measure employed by [6]. Spectral tilt and spectral balance values, syllable durations, and F0 values were scaled and centered using the *scale()* function in R[13] (packages used are: *plyr v1.8.3* and *lme4 v1.1.11*). We discarded syllable durations that exceeded 600 ms as potential alignment errors.

3. Statistical analysis

3.1. Spectral tilt

Figure 1 shows the density plots of stressed (green, solid line) vs. unstressed (blue, dashed line) syllables. The x-axis indicates (scaled and centered) tilt values; the y-axis indicates the likelihood of observing these values. It can be seen that stressed syllables are more likely to exhibit greater tilt values than unstressed syllables: the green, solid line is shifted to the right, relative to the blue, dashed line. This is the exact opposite of what the literature suggests: usually stressed syllables are assumed to have a flatter, less steep slope, which would indicate greater vocal effort. In order to test for statistical significance of the influence of stress on spectral tilt, we employ the following methodology: Following common practice (e.g. [14]), we fit two linear mixed models [15], each predicting the acoustic parameter in question. In one model, we include stress as a fixed factor, in the other, we do not. We include random by-speaker intercepts in both models in order to allow for individual means of spectral tilt for every speaker. We then compare the two models by way of an ANOVA. We consider one model to provide a significantly better fit than the other model if the ANOVA indicates that $p < \alpha$, and if in addition its AIC value is at least 2 points smaller than that of the competing model (cf. [14, 16]). We want to assume an $\alpha = 0.01$ for the present experiment. Since we will conduct this kind of analysis 4 times, once for each parameter, we use Bonferroni correction and set $\alpha = 0.0025$.

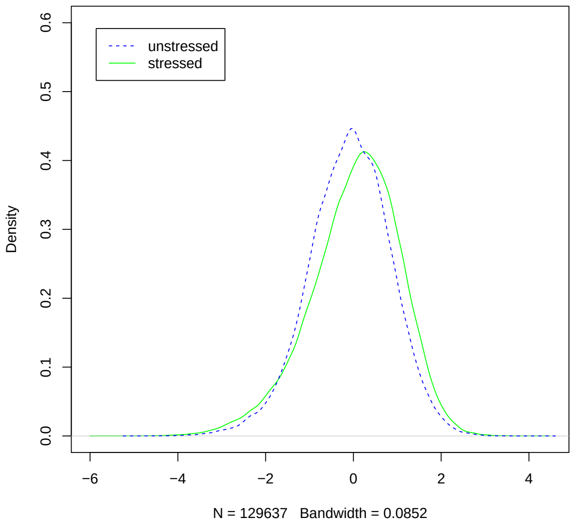

Figure 1: Normalized spectral tilt for stressed and unstressed vowels for all speakers. Stressed vowels have a higher tilt (= steeper slope).

For the present parameter, spectral tilt, we thus compare models (1a) and (1b). The ANOVA determines that the model with stress in (1a) provides a significantly better fit ($\chi^2(1) = 662.72$, $p < 2.2e^{-16}$ and $\Delta AIC = 660$). Therefore we consider the effect of stress on spectral tilt significant.

$$tilt \sim stress + (1|talker) \tag{1a}$$
$$tilt \sim (1|talker) \tag{1b}$$

A potential confound in this analysis could be the impact of function words. It is well known that they are produced in a more reduced way than content words, possibly changing the quality of the vowels we explore for tilt. This way the syllables in function words could be weakened to such an extent that the intensity distribution in the frequency spectrum may become comparable to unstressed syllables (although they are marked as stressed in the lexicon and therefore in our analysis would be counted as stressed syllables). However, excluding all function words based on their part-of-speech tags did not change the graph in any noticeable way, so we omit the corresponding density distribution due to limited space.

3.2. Spectral balance

Figure 2 shows the density plots for the spectral balance of stressed (green, solid line) vs. unstressed (blue, dashed line) syllables. The x-axis indicates (scaled and centered) differences in intensity between bands B1 and B2; the y-axis indicates the likelihood of observing these values. It can be seen that unstressed syllables are more likely to exhibit greater differences between these two bands: the blue, dashed line is shifted to the right, relative to the green, solid line. This confirms the findings by [6]. We use the methodology described above to

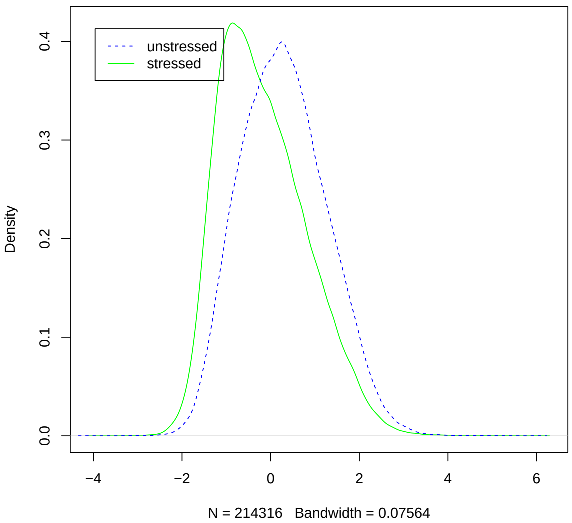

Figure 2: Normalized spectral balance for stressed and unstressed vowels for all speakers. Unstressed vowels (blue line) have greater differences in the intensities in B1 vs. B2 (i.e. unstressed vowels have a steeper spectral slope from B1 to B2).

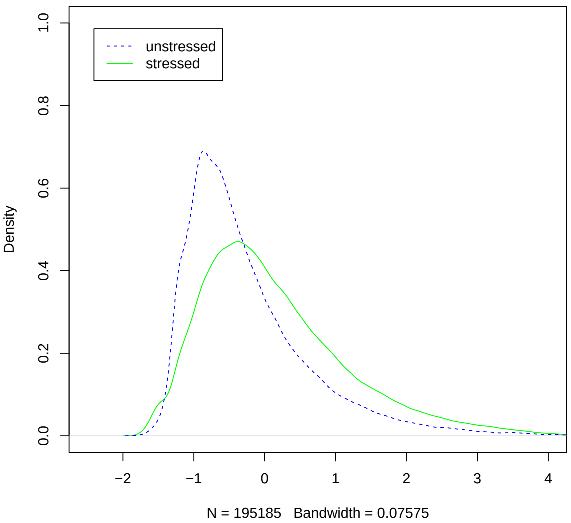

Figure 3: Normalized density plot for duration. Stressed syllables have higher duration and are more widely scattered than unstressed syllables

	Estimate	Std. Error	t value
(Intercept)	0.1727418	0.0050673	34.09
stress	0.0386189	0.0005092	75.85
syl_numphones	0.0436190	0.0002919	149.41

Table 1: Fixed effects of $syl\_dur \sim stress + syl\_numphones + (1|talker)$ with centered $syl\_numphones$.

confirm that the effect is significant ($\chi^2(1) = 17552, p < 2.2e^{-16}, \Delta AIC = 17550$).

3.3. Duration

Regarding duration it is confirmed in Figure 3 that duration is an important cue for stress: Durations of stressed syllables (green, solid line) are shifted to the right compared to unstressed syllables (blue, dashed line). Comparing models with and without stress by an ANOVA confirms that stress affects syllable duration significantly ($\chi^2(1) = 6070.6, p < 2.2e^{-16}, \Delta AIC = 6069$). To make sure that the longer duration of stressed syllables is not simply an effect of different numbers of phones we fit a third model in which we included the (centered) number of phones as an additional fixed effect. This model was found to provide an even better fit ($\chi^2(1) = 21562, p < 2.2e^{-16}, \Delta AIC = 21559$).

The coefficients of that model are indicated in Table 1. They show that indeed the number of phones in a syllable is correlated with the duration of a word: duration increases by about 43 ms for every additional phone. The coefficient for stress of 0.0386 indicates that in addition, and independently of the number of phones, stressed syllables are approx. 39 ms longer on average than unstressed syllables. For the model without number of phones (not shown here), we had obtained a very similar coefficient of 0.0412, i.e. a difference of approx. 41 ms between stressed and unstressed syllables, thus we can conclude that the effect of stress on the duration is preserved even when integrating number of phones as another explaining factor. Similarly, removing function words did not affect the general results.

3.4. Pitch

Looking at the density plots for the pitch parameter we can see that the differences will hardly be significant (see Figure 4). In fact stressed and unstressed syllables have an extraordinarily similar distribution. There is almost no difference between the blue and the green line. An ANOVA using *stress* as a fixed effect and *talker* as a random effect shows that the difference between the models with and without stress is indeed not significant.

4. Discussion & Conclusion

We examined several acoustic parameters that have been suggested to be employed in marking word stress in speech production: For duration, we could fully confirm the wide-spread claim that duration is an important correlate of word stress. The effect was significant, and clearly visible in the density distributions of stressed vs. unstressed syllables.

Regarding spectral tilt, we specifically tried to reproduce the finding by [6] that stressed syllables have a lower tilt than unstressed syllables. On German data the finding could only be confirmed partially. When calculating what we called spectral *balance* by dividing the spectrum into frequency bands and comparing intensities in the lowest two bands, i.e. between 0 to 500 Hz and between 500 and 1,000 Hz, we could confirm that there is a less steep spectral slope for stressed vowels than for unstressed vowels. When calculating what we called spectral *tilt* by linear regression over the frequency spectrum, we found that stressed syllables have in fact a higher spectral tilt than unstressed syllables, which is a very unexpected finding, exactly contradicting the findings by [6]. Several factors were different in our experiment. While [6] had a more artificial setup

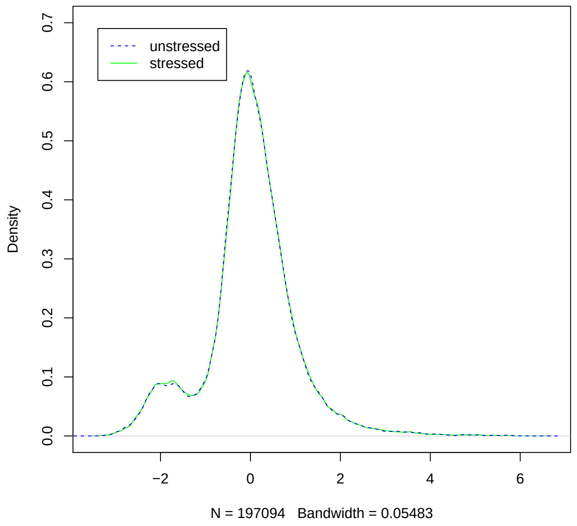

Figure 4: Normalized F_0 for stressed and unstressed vowels for all speakers with function words. Pitch for stressed and unstressed vowels is almost identical.

with participants producing very specific words, we analyzed natural speech, trying to create conditions that are as close to a real world setting as possible. Such scenarios cause a lot of uncontrolled variation in the data, which may be reflected in the seemingly inconsistent results. Also, we would like to point out that spectral balance as defined above looks only at the spectrum between 0 and 1,000 Hz. Preliminary results (not shown here) indicate that the differences are completely lost, and partially reversed in the higher frequency bands. Thus the contradictory results may arise simply as a consequence of looking at different ranges of frequency. This would imply that the wide-spread claim that unstressed vowels have a steeper spectral slope holds only for the lower frequencies.

Finally, while many authors consider pitch to be an important correlate of word stress, we did not find a significant effect. This might be due to the fact that we measured pitch in the middle of the vowel while prevalent literature [10] looks at the maximum F_0 of the vowel. Another possible reason is that other studies can control for pitch accent, due to their experimental design, while in our study we did not have information about the location of pitch accents.

5. Bibliographie

[1] D. C. Walker, *French sound structure*. University of Calgary Press, 2001, vol. 1.

[2] M. J. W. Yip, “The tonal phonology of chinese,” Dissertation, Massachusetts Institute of Technology. Dept. of Linguistics and Philosophy, 1980.

[3] D. B. Fry, “Duration and intensity as physical correlates of linguistic stress,” *Journal of the Acoustical Society of America*, vol. 27, pp. 765–768, 1955.

[4] ——, “The dependence of stress judgements on vowel formant structure,” *Language and Speech*, vol. 1, pp. 126–152, 1958.

[5] ——, “The Dependence of Stress Judgements on Vowel Formant Structure,” in *Proceedings of the Fifth International Congress of Phonetic Science*, 1965, pp. 306–311.

[6] A. M. Sluijter and V. J. Van Heuven, “Spectral balance as an acoustic correlate of linguistic stress,” *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[7] A. M. Sluijter, V. J. van Heuven, and J. J. A. Pacilly, “Spectral balance as a cue in the perception of linguistic stress,” *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 503–513, 1997.

[8] N. Campbell and M. Beckman, “Stress, prominence, and spectral tilt,” in *Intonation: Theory, Models and Applications: Proceedings of an ESCA Workshop*, A. Botinis, G. Kouroupetroglou, and G. Carayiannis, Eds. ESCA ETRW, 1997.

[9] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, “Loudness predicts prominence: Fundamental frequency lends little,” *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1038–1054, 2005.

[10] A. O. Okobi, “Acoustic correlates of word stress in american english,” Dissertation, Cornell University, 2006.

[11] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” 2015, September 2015, Edition: 5.4.22. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[12] A. Schweitzer and N. Lewandowski, “Social Factors in Convergence of F1 and F2 in Spontaneous Speech,” in *Proceedings of the 10th International Seminar on Speech Production, Cologne*, 2014.

[13] R Core Team, “R: A language and environment for statistical computing,” R Foundation for Statistical Computing, Vienna, Austria, 2015, Edition: 3.2.3. [Online]. Available: https://www.R-project.org/

[14] B. Winter, “Linear models and linear mixed effects models in R with linguistic applications,” *arXiv preprint arXiv:1308.5499*, 2013.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[16] K. P. Burnham and D. Anderson, “Model selection and multimodel inference,” *A Practical information-theoretic approach. Springer*, 2003.