Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

# Beware of the individual: Evaluating prominence perception in spontaneous speech

*Anna Bruggeman*[1], *Leonie Schade*[1], *Marcin Włodarczak*[2], *Petra Wagner*[1]

[1]Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University
[2]Stockholm University

{anna.bruggeman,petra.wagner}@uni-bielefeld.de, marcin.wlodarczak@ling.su.se

## Abstract

Much of the existing research on prominence perception has focused on read speech in American English and German. The present paper presents two experiments that build on and extend insights from these studies in two ways. Firstly, we elicit prominence judgments on spontaneous speech. Secondly, we investigate gradient rather than binary prominence judgments by introducing a finger tapping task. We then provide a within-participant comparison of gradient prominence results with binary prominence judgments to evaluate their correspondence.

Our results show that participants exhibit different success rates in tapping the prominence pattern of spontaneous data, but generally tapping results correlate well with binary prominence judgments within individuals. Random forest analyses of the acoustic parameters involved show that pitch accentuation and duration play important roles in both binary judgments and prominence tapping patterns. We can also confirm earlier findings from read speech that differences exist between participants in the relative importance rankings of various signal and systematic properties.

**Index Terms**: prominence, individual differences, German

## 1. Introduction

Prominence is a complex feature of speech; it is relative, context-dependent, and its perception may be based on any combination of semantic, pragmatic, syntactic or prosodic properties of the signal. Prominence judgments on speech have been adressed in detail for American English [1, 2, 3, 4] and in German [5, 6], as well as in French and Spanish [4] and Czech [7]. A central goal in this work is to understand the task division between different properties of speech in explaining prominence perception. In the more recent studies [5, 6] it was noted that the cues attended to are highly listener-dependent, with some participants being more oriented towards the acoustic signal, and others more attuned to structural factors such as part-of-speech. At this point it is not fully clear whether further distinctions can be made within these groups, although the expectation would be that there likely are further differences among 'acoustic' listeners (cf. [8] on speaker-specific patterns in German pitch accent perception).

Thus far, a typical prominence perception experiment involves listeners simultaneously listening and reading speech materials, and subsequently underlining or highlighting words perceived as prominent, a so-called Rapid Prosody Transcription (RPT) task. The resulting binary prominence ratings are then aggregated across participants. Following [6], who employed a drumming methodology, the present paper will look at prominence perception in a more gradient way, by employing an adapted technique that requires participants to finger-tap a perceived prominence pattern in syllable tact, after hearing the stimulus. This removes any potential effects of participants' simultaneous reading and listening to stimuli, and allows for the assessment of gradient prominence judgments at the individual level. A second aspect the present experiments contribute is the extension of research on prominence perception (in German) to spontaneous speech. We will assess whether the clear correlation between signal and prominence judgments reported for read speech also holds for spontaneous speech containing non-canonical accent realisation, hesitations, reduction phenomena etc. Finally, the tasks employed here involve tapping to syllables (rather than words), which allows us to zoom in on the correspondence between prominence judgments and acoustic properties of that particular syllable.

The first experiment served as a pilot to test the tapping methodology and to gain a first insight into the extent of between-participant variation. The second experiment optimised the tapping task, and links tapping and RPT performance within individuals.

## 2. Exp. 1: Tapping to inter-pausal units

### 2.1. Methodology

#### *2.1.1. Speech materials*

Speech materials to be judged for prominence were taken from the IMS GECO corpus [9], containing spontaneous, free dialogue between female speakers who did not know each other prior to recording. From this corpus, three speakers were selected that exhibited few regional characteristics in their speech. Inter-Pausal Units (IPUs) were automatically extracted and a subset of these ranging in length from 6 to 15 syllables was used for prominence tapping.

#### *2.1.2. Participants and task*

Four native speakers of standard German (western and northern) served as participants. They were presented auditorily, through headphones, with single IPUs in isolation. Upon listening to an IPU, participants were asked to indicate their perceived prominence pattern by means of finger tapping on a Sensel drumpad [10], representing the strength of each syllable with a corresponding tap. Tapping occurred after listening to the entire IPU. Auditory feedback in the form of a drumbeat sound was given during the tapping, and participants were allowed to retap up to 3 items during the course of the experiment. Each participant provided tapping responses to 90 IPUS: 30 IPUs that were presented to all, and a further 60 unique IPUs that differed for each participant. The total number of responses was N=270. A training phase of 15 IPUs with lengths of 6 or 7 syllables preceded the actual experiment.


10.21437/SpeechProsody.2022-55

#### 2.1.3. Data processing and statistics

Tapping responses on the Sensel pad are recorded in terms of both impact force and duration of contact, and were imported into the software Ableton Live [11]. For this experiment, only impact force was used (henceforth tap force or TF). Absolute TF values ranged from 13 to 127 and were log-transformed and subsequently normalised (z-scored) for each IPU on a by-participant basis. The following acoustic properties of the signal were extracted in Praat [12], in each case for both vowels and syllables: DUR (duration in ms.) $F_0$ (max, mean, range, in ST), INT (intensity in dB). Additionally, voice quality measures were taken from vocalic portions of the signal only: CPPS (calculation as in [13]), H1H2 (difference in amplitude between first and second formant, in dB) and spectral SLOPE (difference in amplitude between energy bands 50-1000 Hz and 1-5 kHz, in dB). All data was z-scored per IPU, and the two duration measures were first log-transformed.Given the known correlations between many of the above predictors, the analysis presented here is based on random forests (RFs) allowing for an assessment of each predictor's contribution to TF values. RFs were constructed for each participant by means of the cforest function in the R package *party* [14, 15, 16]. Datapoints were assigned randomly either to a training set containing 70% of data, or a validation set containing the remaining 30%. To assess goodness of fit, $R^2$ values were then calculated on the predictions vs. actual TF values for each validation dataset. Means will be given for N=100 models per participant.

### 2.2. Results

#### 2.2.1. Tapping success

Tapping success was defined in terms of a number of taps that matched the number of syllables in a given IPU. With increasing IPU length, tapping success decreased, to the point where IPUs with 15 syllables were only correctly tapped to in 36% of all cases (N=36). In contrast, for the shortest IPU length of 6 syllables (N=36), the mean success rate was 80%. In addition to this, there were considerable inter-speaker differences in overall success, with a range of 36% – 64% correct (N=90 per participant). These findings indicate that the task was quite challenging overall, with added difficulty for longer IPU length. Post-experiment evaluation suggested that the difficulty was related more to issues with recall of the number of syllables, and sometimes of the content non-coherent speech units, than to the tapping task itself. Further support for this interpretation comes from [6], using read speech, where success rate for drumming with drumsticks was around 95%. We will address these considerations in the second experiment.

#### 2.2.2. Acoustic correlates of prominence ratings

Details of explained variance ($R^2$) and variable importance rankings are given in Table 1:

Table 1: *Exp 1: Top 3 variables in importance ranking*

| | Var 1 | Var 2 | Var 3 | $R^2$ |
|---|---|---|---|---|
| P1 | dur vowel | dur syll | $F_0$ range | 0.11 |
| P2 | dur syll | slope | dur vowel | 0.05 |
| P3 | slope | intens | $F_0$ max | 0.02 |
| P4 | dur syll | $F_0$ max | $F_0$ mean | 0.03 |

The two most common predictors included DUR (vowel, syllable)and $F_0$ (mean, max and range). Individual differences nevertheless exist,with two participants paying attention to a measure of voice quality, SLOPE, and one to intensity (INTENS). Explained variance for the individual models is however quite low, which indicates that participants' prominence judgments were likely influenced to a large extent by factors not considered here.

### 2.3. Interim summary

The results of Experiment 1 showed that the experiment was somewhat challenging, in terms of providing matching number of taps to the number of syllables heard in an utterance. Beyond this, inter-participant differences were visible at the level of cue importance, which is in line with other recent findings on prominence perception in German [5, 6]. The next experiment aims to reduce the difficulty with the production of the expected number of taps. We also extend the task so that participants both tap to, and provide RPT (underlining) judgments on the same data.

## 3. Exp. 2: Tapping to intonational units

### 3.1. Methodology

#### 3.1.1. Speech materials

Speech materials came from the same corpus and the same three speakers as in experiment 1. For this second experiment, a random selection of N=53 phrases with length 8 to 11 syllables was extracted. Due to the low success rate for recall in experiment 1, care was taken to select well-formed intonational units, which were typically Intonational Phrases (IPs) and in some cases intermediate phrases (ips). The nature of spontaneous speech leads, in some cases, to ambiguity with respect to perceived syllabification; some words and collocations were rendered with fewer phonetic than lexical syllables, e.g. [aʊ̯s.t͡siːn] for *ausziehen*, and [ha.pɪç] for *habe ich*. Divergences from standard syllabification were coded for and will be taken into account. Additionally, several phrases also contained disfluencies or fillers such as *m-m* and *ahh*, in which cases such syllables were coded as non-lexical. Out of the N=53 target sentences, N=12 had potentially ambiguous syllabification and N=7 contained a non-lexical element. Participants were instructed to tap to the number of actual syllables they heard, including fillers and the like.

#### 3.1.2. Participants and task

12 native German-speaking participants (5m, 7f) were recruited among university students in Bielefeld. Most, though not all, participants were from northern or western Germany, many had some linguistic training, and many had musical experience.

Similar to experiment 1, there was a training phase consisting of 16 items ranging in length between 6 and 11 syllables. In the main experimental phase, participants performed the tapping task on the target items (N=53), where they were allowed to ask for up to 3 retakes (as in experiment 1). The order of presentation was counterbalanced across participants, but utterances of a given speaker were always heard in one continuous block.

After the tapping task, participants also performed an underlining task on the same target materials. Following the RPT method laid out in [5, 1], the instructions asked participants to underline whichever syllables they perceived to be prominent ('betont / hervorgehoben / wichtig'). This served to assess, on an individual level, whether different measures of participants' prominence perception yield similar results. The total duration of the experimental session was less than an hour.

#### 3.1.3. Data processing and statistics

Data preprocessing was similar to in experiment 1, with the exception of the implementation of SLOPE: here we used a measure of the amplitude difference (in dB) between frequency bands 0-1 kHz (as opposed to 0-0.5 kHz) and 1-5 kHz. We henceforth refer to it as ALPHA ratio. Acoustic measurements were taken for all N=475 syllables (53 sentences with between 8 and 11 syllables each). Outliers on all measurements were checked and a small number was removed. Information was added for several additional variables that are known to play a role in prominence perception. These include two prosodic factors: phrasal position (PHRASE-POS) (initial, final, or medial), added to account for a possible effect of edge-related enhancement of acoustic parameters; and PA, the presence of a pitch accent (based on visual and auditory checking by the first author). Additional structural factors that were added include NSYLL to take into account the number of syllables in the word; FREQ (log-word frequency from the SUBTLEX database [17]); and POS (part-of-speech) reflecting whether the syllable is part of a function or a content word.

### 3.2. Results

#### 3.2.1. Tapping success

The total number of utterances for which there are tapped responses is N=633 (12 participants x 53 items minus 3 lost datapoints). Overall, there were fewer unsuccessfully tapped items in this experiment than in experiment 1. 69% of items tapped to had a number of taps which matched the number of syllables (range between participants: 60% to 85%). Among these, sentences with ambiguous syllabification due to reduction did not appear much more difficult; these had the correct number of taps in 60% of cases. We used a single reference syllabification (the first impression of the first author), so this reflects a conservative estimate of success. Similarly, sentences with disfluencies were correctly tapped to in 67% of cases. As the lower end of the success rate range in this experiment is similar to the higher end of the success rate in experiment 1, the changes in design appear successful in terms of making the task more straightforward.

#### 3.2.2. Tapping vs. RPT

Tap Force (TF) and Tap Duration (TD) for each syllable were correlated with the same participant's prominence judgment in the RPT underlining task. Two logistic regression models predicting RPT prominence (underlined: 'prom', not underlined: 'not prom') were fitted for each participant, one based on TF and one on TD (log-transformed and normalised). Figure 1 shows the difference in estimates: most participants used, on average, both longer taps and greater force for syllables that they subsequently judged prominent in the RPT task. Significant differences between prominence categories are indicated with green lines ($p < 0.05$ for the main effect of TD or TF). One participant, P8, did not differentiate their underlined categories in their tapping. P1 only significantly differentiated between the two categories by means of force, not duration, and the reverse holds true for P3 and P11. In addition, there were great individual differences in the extent to which both tap measures were correlated, ranging from r=0 to r=0.66, mean r=0.33. This shows that participants may employ different strategies in tapping prominence patterns. We take these results to indicate that nearly all participants are able to perform the tapping task in this format, and thus that they can translate perceived strength relationships between syllables in a sentence to a different modality, in this case finger tapping. Yet it is clear that there is no one single strategy employed by all, something we will keep in mind for further analysis.

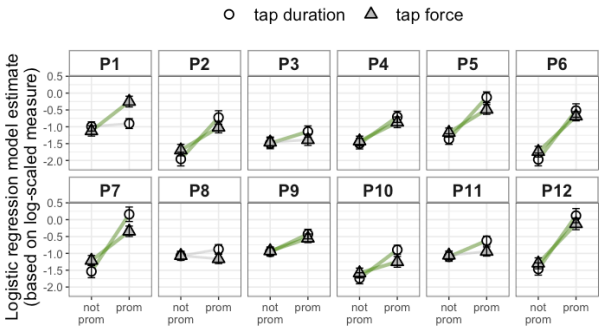


Figure 1: *Model estimates for TD and TF. Green lines indicate significant tap differences for syllables judged prominent/not prominent in the RPT task.*

Finally, participants were reasonably aligned in terms of their prominence underlining judgments, with a Fleiss-Kappa value of 0.51 (the same as in the word-based RPT in [5]).

#### 3.2.3. Acoustic correlates of prominence ratings

RF regressions were run for Tap Force for most participants, and for Tap Duration for participants P3 and P11. In first instance, forests were constructed for individual participants and based on acoustic properties only, specifically vowel measures INT, $F_0$ (mean/max/range) CPPS, ALPHA, H1H2, and DUR of both vowels and syllables. Table 2 gives the variable ranking for each participant (negative variable importance excluded in two cases) and the corresponding $R^2$, ranging from 0.01 to 0.13. As in experiment 1, the numbers are based on 100 forests per participant.

Table 2: *Exp 2: Top 3 variables in importance ranking*

| | Var 1 | Var 2 | Var 3 | $R^2$ |
|---|---|---|---|---|
| P1 | dur vowel | dur syll | alpha | 0.07 |
| P2 | dur syll | $F_0$ max | dur vowel | 0.04 |
| P3 | dur vowel | alpha | dur syll | 0.06 |
| P4 | intens | dur syll | dur vowel | 0.06 |
| P5 | dur syll | dur vowel | cpps | 0.06 |
| P6 | dur syll | alpha | dur vowel | 0.10 |
| P7 | dur syll | cpps | $F_0$ range | 0.03 |
| P8 | cpps | $F_0$ range | | 0.01 |
| P9 | dur syll | cpps | | 0.01 |
| P10 | dur syll | alpha | intens | 0.01 |
| P11 | dur syll | dur vowel | $F_0$ mean | 0.13 |
| P12 | cpps | dur vowel | $F_0$ max | 0.01 |

The same acoustic-only analysis was then performed on the full dataset, collapsing data from all participants. This yielded a mean $R^2$ of 0.07 and the variable importance ranking in Figure 2. On the whole, as across participants, the strongest predictors include durational and voice quality measures.

Some of the variability in prominence tapping could likely be explained by other factors that are known to contribute to prominence perception. We thus constructed forests with a combination of acoustic predictors (the 4 highest ranking from the

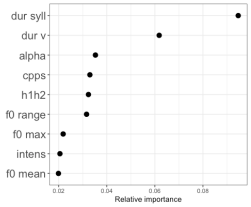


Figure 2: *Acoustic forests*

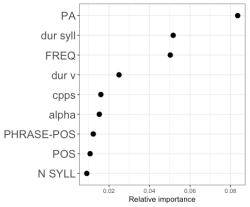


Figure 3: *Mixed forests*

acoustic model: DUR syllable/vowel, ALPHA and CPPS) and 5 additional predictors mentioned in section 3.1.3: PA, POS, NSYLL, FREQ and PHRASE-POS. $R^2$ for these forests did not increase (i.e. was also 0.07), and the resulting predictor ranking is shown in Figure 3, with non-acoustic predictors capitalised. For forests run on each participant's data individually, the range was $R^2$ 0.01 – 0.30 (mean 0.11), with the predictor PA far outranking the other predictors for 3 participants.

# 4. Discussion

## 4.1. Prominence tapping and spontaneous speech

At first glance, the explanatory value of our chosen predictors seems rather low, especially in the case of models with acoustic predictors only. Higher $R^2$ values of up to 0.30 for individual participants were reached only in cases where additional factors including pitch accentuation and word frequency were added. For comparison, an $R^2$=0.71 or 71% explained variance was reported by [5] on binary RPT prominence judgments.

There are a few possible explanations for the present lower $R^2$ values. One concerns the use of aggregated data. To be able to compare directly, we subjected our own RPT judgments to the aggregate analysis from [5] (code and package *ranger* [18]). RFs are thus run on one mean prominence value per syllable. Using our above set of 9 mixed predictors, which largely overlap with those in [5], RFs on averages yielded a much higher $R^2$ of 0.62. We however did not include the second- and thirdmost important predictors from [5], namely accent type and accent position. It is encouraging that RFs nevertheless reached 62% explained variance on the present data, especially given that the two datasets also reflect a difference in speech style. Using the same method with acoustic predictors alone, the explained variance for our averaged RPT data was 58%.

In contrast, mixed forests predicting averaged TD/TF values yielded lower explained variance, around 22% for mixed, and 13% for acoustic-only forests. This direct comparison shows that, when looking at averages, RPT judgments appear to be more strongly linked to known predictors of prominence than tapping patterns. It also shows that the choice of RF implementation, and averaging across participants, yields different results. In sum, considering the comparability of averaged results across studies, the lower explained variance for individuals' TD/TF might also turn out to be realistic.

One of our further questions was whether syllabic tapping would predispose participants towards a greater reliance on acoustic features. The present setup with spontaneous speech showed that, despite considerable reliance on purely acoustic properties of the signal (e.g. $R^2$=0.58 for RPT judgments), models including additional predictors generally fared better. This holds irrespective of whether the task involved tapping or underlining. Pitch accentuation and word frequency are important predictors, and suggest that even in a syllabic motor task, systematic and top-down knowledge of language continues to play a critical role.

Finally, the relative importance of the various acoustic parameters was as expected. Duration was consistently among the most important predictors of both RPT judgments and tapping, which is in line with its known role in signalling prominence in Germanic languages generally [19, 20, 21, 22]. $f_0$, too, is known to be an important contributor to prominence perception, as reflected in its role in pitch accentuation. Yet while pitch accentuation was a primary predictor, purely signal-based $F_0$ measurements did not feature high in the importance rankings. This could be due to the nature of the speech stimuli, with spontaneous speech potentially exhibiting a greater degree of $F_0$ modulation for non-prominence lending purposes than controlled, read speech. It is conceivable that participants would then rely to a greater extent on top-down knowledge, notably word frequency in the present results, and on other acoustic cues including duration and voice quality variation (Figure 3). For now this remains speculative, although similar ideas have also been expressed in [23]; future work could seek to investigate the role of voice quality and potential trading relations with other cues further.

## 4.2. Individual differences in prominence judgments

Firstly, there were some differences in how individuals' tapping behaviour correlated with their prominence judgments in the RPT. For 1 participant, the two did not correlate at all, and 3 more relied on either TF (tap force) or TD (tap duration) alone. Secondly, acoustic-only RFs did not account very well for the variability in TD/TF for any of the participants. For specific individuals, RFs that included non-acoustic factors accounted for somewhat more variance (over 10% for 4 participants). As expected, we also observed that individuals exhibited differences in what aspects of the signal (or linguistic system) they attended to in making these prominence judgments. Among acoustic predictors, duration and voice quality measures took precedence, but some individuals attended to $F_0$ as well. Finally, considerable variation was also found between participants in the extent to which prominence judgments from the RPT and pitch accentuation correlated (Cramér's V: 0.36–0.75). Even on the lower end of this range there is nevertheless some association between the two, which is in line with [4] and [5].

# 5. Conclusions

The experiments reported here have shed light on the acoustic properties of spontaneous speech that contribute to prominence perception in German, and individual differences therein. Recent studies investigating German read speech have reported considerable individual differences in terms of the aspects of speech attended to in judging prominence [6, 5], and we confirmed the existence of similar differences for spontaneous speech. Based on various random forest analyses, we found that variation in gradient prominence tapping behaviour was accounted for less well than averaged binary RPT judgments.

# 6. Acknowledgements

This research is supported by the Swedish Research Council grant *Prosodic functions of voice quality dynamics* (2019-02932) to the third author. The authors are grateful to Bogdan Ludusan and Jana Wiechmann for feedback on data and the setup details.

# 7. References


[1] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, pp. 425–452, 2010.

[2] J. Cole, Y. Mo, and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 1141–1177, 2010.

[3] J. Cole and S. Shattuck-Hufnagel, "New methods for prosodic transcription: Capturing variability as a source of information," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 7, no. 1, 2016.

[4] J. Cole, J. I. Hualde, C. L. Smith, C. Eager, T. Mahrt, and R. N. de Souza, "Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish," *Journal of Phonetics*, vol. 75, pp. 113–147, 2019.

[5] S. Baumann and B. Winter, "What makes a word prominent? predicting untrained German listeners' perceptual judgments," *Journal of Phonetics*, vol. 70, pp. 20–38, 2018.

[6] P. Wagner, A. Ćwiek, and B. Samlowski, "Exploiting the speech-gesture link to capture fine-grained prosodic prominence impressions and listening strategies," *Journal of Phonetics*, vol. 76, 2019.

[7] L. Weingartová and J. Volín, "Short-term spectral slope measures and their sensitivity to speaker, vowel identity and prominence," *Akustické listy*, vol. 20, no. 1, pp. 5–12, 2014.

[8] M. Grice, S. Ritter, H. Niemann, and T. B. Roettger, "Integrating the discreteness and continuity of intonational categories," *Journal of Phonetics*, vol. 64, no. 1, pp. 90–107, 2017.

[9] A. Schweitzer, N. Lewandowski, D. Duran, and G. Dogil, "Attention, please! expanding the GECO database," in *Proceedings of International Congress of Phonetic Sciences XVIII*, 2015.

[10] Sensel Inc., "Sensel morph," Pressure sensor tablet.

[11] Ableton AG, "Ableton live 10 standard," Software.

[12] P. Boersma and D. Weenink, "Praat," Software, 2015, version 6.0.22.

[13] C. R. Watts, S. N. Awan, and Y. Maryn, "A comparison of cepstral peak prominence measures from two acoustic analysis programs," *Journal of Voice*, vol. 31, no. 3, pp. 387e1–387e10, 2017.

[14] T. Hothorn, P. Buehlmann, S. Dudoit, A. Molinaro, and M. Van Der Laan, "Survival ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, 2006.

[15] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 25, 2007.

[16] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 307, 2008.

[17] M. Brysbaert, M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte, and A. Böhl, "The word frequency effect," *Experimental psychology*, vol. 58, pp. 412–424, 2011.

[18] M. N. Wright and A. Ziegler, "ranger: A fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, vol. 77, no. 1, pp. 1–17, 2017.

[19] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 765–768, 1955.

[20] ——, "Experiments in the perception of stress," *Language and Speech*, vol. 1, no. 2, pp. 126–152, 1958.

[21] T. Cambier-Langeveld, *Temporal marking of accents and boundaries*, ser. LOT series. The Hague: Holland Academic Graphics, 2000.

[22] K. J. Kohler, "The perception of lexical stress in German: effects of segmental duration and vowel quality in different prosodic patterns," *Phonetica*, vol. 69, no. 1-2, pp. 68–93, 2012.

[23] A. N. Chasaide, I. Yanushevskaya, J. Kane, and C. Gobl, "The voice prominence hypothesis: the interplay of F0 and voice source features in accentuation," in *Proceedings of Interspeech 2022*, 2013, pp. 3527–3531.