The Second International Conference on Tone and Intonation
18-20 November 2023, Singapore, Singapore

The perception of German wh-phrase-final intonation

Heiko Seeliger[1], *Anne Lützeler*[1], *Constantijn Kaland*[2]

[1]Department of German Language and Literature I, University of Cologne, Germany
[2]Department of Linguistics, University of Cologne, Germany

heiko.seeliger@uni-koeln.de, anne.luetzeler@uni-koeln.de, ckaland@uni-koeln.de

Abstract

This article follows up on a previous study that performed cluster analysis of F0 contours resulting from a production study on German *wh*-questions and *wh*-exclamatives. We conducted a perception study using hummed versions of contours from two of the clusters, i.e. speech from which no segments are intelligible anymore, while preserving prosodic cues such as intensity, duration and F0 entirely. The goal was to assess whether contours that clustered together were also *perceived* to be similar by native listeners of German. Overall, the results indicate that listeners were able to carry out this fairly abstract task successfully. In the details, the results show a somewhat complex picture: While contours from different clusters were indeed judged to be less perceptually similar than contours from the same clusters, ratings were also influenced by contour duration, speaker gender and/or F0 register. The results thus indicate that the perceived similarity of abstracted F0 contours is sensitive to essentially all aspects of prosody.

Index Terms: boundary tones, perception, cluster analysis, wh-questions, German

1. Introduction

A central question in intonation research concerns its alleged categorical nature. Traditional (autosegmental-metrical) approaches are strongly based on the idea that intonation is a composite of H and L tones, either as (pitch) accents or boundary tones ([1]). Under this view, an intonation contour is a buildup of (combinations of) H and L tones from an inventory. Categories are assumed at the level of this inventory; i.e. a language is assumed to have a limited number of pitch accents and boundary tones with which it constructs all its intonation contours, which in turn derive their meaning from their (composed) shape. This view has been challenged by many psycholinguistic studies (e.g. [2]) and by corpus studies of spontaneous speech (e.g. [3]) showing that speakers and listeners do not maintain a strict categorical separation of intonational form-meaning relationships. On the other hand, studies keep showing that small shape differences in F0 can lead to meaningful differences (e.g. [4]). Thus, to what extent intonation contours are indeed categorical remains an ongoing research question.

Recently, cluster analysis has gained popularity as a methodological approach to study F0 contours and their meaning. While older applications of cluster analysis in prosodic research focused on aggregated measures, such as e.g. per-syllable mean pitch across an utterance (e.g. [5]), recent studies have started applying cluster analysis directly to F0 contours (e.g. [6], [7], [8], [9]). This allows the investigation of more minute differences, which might well become undetectable through aggregation. The results of applying cluster analysis directly to F0 contours have been promising so far, in that it is usually the case that phonologically meaningful clusters emerge. Even in the cases where a phonologically expected contrast does *not* emerge in a cluster analysis, a close analysis of the contours involved can help shed light on phonetic and/or contextual variation within (presumed) phonological categories.

However, cluster analysis on F0 contours requires perceptual validation. If a cluster that emerges from a cluster analysis is not actually perceptually distinct from members of another cluster, this might indicate that the number of clusters assumed was too high, and/or that the cluster analysis was sensitive to small, inaudible differences. The relevance of doing perceptual evaluations on the outcomes of cluster analyses on F0 contours has been shown in a recent study ([10]). In that study, the perception of F0 contour differences was tested with listeners of German and Papuan Malay, and compared to acoustically measured representations of the same contours (in ERB, standardized, OMe rescaled [11] or as first derivative) and their differences (Euclidean, Pearson, Dynamic Time Warping). The presented contours were originally taken from spontaneously produced Papuan Malay speech, then stylized and hummed, such that listeners were only presented with F0 differences between contours (no segmental content, no acoustic differences otherwise). The results showed that both listener groups perceived F0 differences in a highly similar way. As for the comparison with the measured contour representation, this study showed that contours represented using their first derivative with dynamic time warping to quantify their differences showed the highest correlation with the F0 differences perceived by humans. Nevertheless, correlations were moderate, still leaving a considerable amount of improvement to be bridged between carefully chosen numerical representations and highly controlled human perception.

The current study uses a similar paradigm in order to test two clusters from the cluster analysis described in [12] in a perception study. Section 2 gives background information on that analysis, as well as the theoretical area of interest. Section 3 presents the perception experiment and its results. Section 4 concludes.

2. Cluster analysis on F0 contours from German *wh*-questions and *wh*-exclamatives

The present article builds on [12], who performed a cluster analysis of F0 contours from a production study ([13]). Target sentences were German *wh*-questions and *wh*-exclamatives; see example (1) for a sample item. The production study manipulated the information structure of target sentences in terms of contrast and givenness, which led to variation in nuclear accent location and boundary tone choice. The cluster analysis of F0 contours

10.21437/TAI.2023-3

described in [12] was initially motivated by the impression that there were comparatively many medium-high, level plateaus in utterance-final position. Cluster analysis was chosen as a tool to semi-automatically classify the whole data set, so that the plateaus could be isolated and studied further.

(1) a. Wo die schon überall
where she-DEM *already everywhere*
Germanen erforscht hat!
Germanic.tribes researched has
'The places where she has already researched Germanic tribes!'

b. Weißt du zufällig, [= matrix polar question]
wo die schon überall Germanen erforscht hat?
'Do you happen to know where she has researched Germanic tribes already?'

[12] combined two separate cluster analyses: One on the whole utterance, with 20 evenly spaced F0 points; and one on the final two syllables, with 10 evenly spaced F0 points. In this way, the combined analysis classified the entire contour, while simultaneously giving extra weight to utterance-final pitch. We used hierarchical clustering on Euclidean distances (see [12] for more detail on the original cluster analysis, and [14] and [15] for more background on contour clustering in general and the R [16] app developed by the third author of this article).

The relevant contours in the two clusters of interest consisted of medium-high plateaus, some of which featured steep, late falls. In terms of GToBI labels [17], H-L% suggests itself for the late falls (but is not part of the GToBI system of boundary tones). For the level plateaus, H-% or !H-% might be appropriate, although the former has mostly been described as a continuation rise and the latter features most prominently as the final tone of the calling contour (cf. [18]) – neither of which appear to be fitting characterizations of the relevant contours. We will refer to the two cluster as Fall and Level for short in the remainder of this article.

3. Perception experiment

In order to find out to which extent the two clusters of interest, which have both commonalities and differences, also exhibit similarities in perception, we designed a perception study in which we elicited ratings of perceived similarity for contour pairs.

3.1. Methods & design

We selected the 10 most typical contours from each of the two clusters, where 'typical' meant the smallest differences to the average contour within each cluster. For the Fall cluster, we selected the 5 most typical steeply falling contours and the 5 most typical shallowly falling contours. We selected only questions (which constituted the majority of each cluster, although both clusters contained both sentence types), i.e. structures like the embedded, underlined clause in (1b). We originally intended to only compare contours between clusters, and to compare every possible contour pair, for a total of 100 comparisons to be rated by each participant. Due to an oversight, however, we also compared some contours *within* clusters (see Table 1 for a breakdown of which comparisons were made). While this was not an intended feature of the design, the within-cluster comparisons can serve as a kind of control condition.

Fig. 1 shows the 20 contours that were used in the perception study. Several cluster-inherent features are already visible here: Fall contours tended to be longer than Level contours; there was only one contour spoken by a female speaker in the Level cluster, while the Fall cluster was more evenly mixed; and the steep, late falls in the Fall cluster were produced only by female speakers (all but one by the same speaker). Note, however, that also the Fall contours produced by male speakers exhibit a downward drift of F0 that is not present in the Level contours.

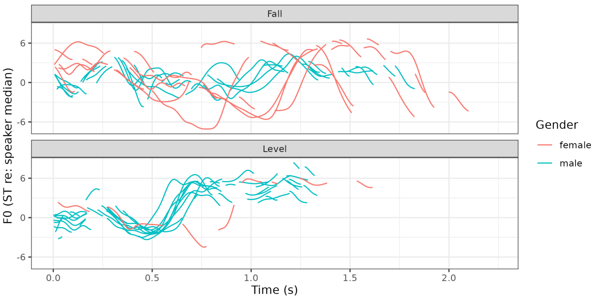

Figure 1: *F0 contours used in the perception study, split up by original cluster and colored by speaker gender*

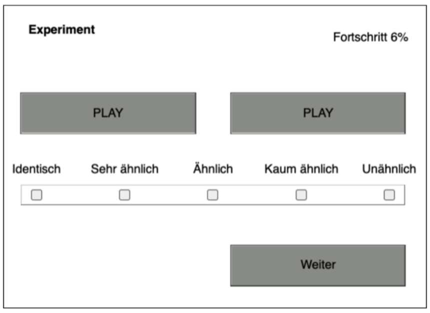

Figure 2: *A screenshot of the experimental setup*

Because the clusters systematically differed in the text that was spoken in the original recordings, we decided to preprocess the contours, so that only F0 and duration were available as cues. The processing was performed in Praat [19], by converting the sound to a Pitch object, interpolating and smoothing it, and then converting it back into a hummed sound. The final stimuli had F0 and static formants, but no segmental differences. We did not normalize the durations of the contours, because different durations appeared to be a feature of the clusters (despite the original cluster analysis using normalized duration).

21 listeners participated in the perception study (7 male, 14 female; mean age = 28.5, age range = 20–63). They gave informed consent and were not paid for their participation. Completing the rating task took on average 25 minutes. The experiment was run using OpenSesame [20] and a custom Python script. Participants' task was to judge how similar two contours sounded. The instructions specified that participants were to pay attention only to the overall sound of the contour and not to differences between speakers and/or genders. Ratings were given on a 5 point rating scale, with points labeled *identisch – sehr ähnlich – ähnlich – kaum ähnlich – unähnlich* (German for "identical", "very similar", "similar", "hardly similar", "dissimilar"). We mapped these ratings to numbers from 5 to 1 for the statistical analysis. Contours were presented in pairs, with the order of contour pairs randomized for each participant. Participants could replay stimuli as often as desired. Fig. 2 illustrates

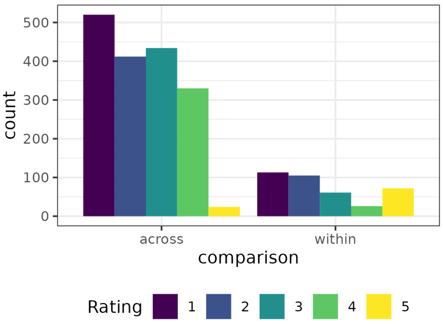

Figure 3: *Barplot of similarity ratings*

the experimental setup as seen by participants.

3.2. Results

We collected 2097 ratings from participants (21 participants × 10*10 comparisons, with three cases of missing data). Table 1 shows how these ratings are distributed across clusters and speaker genders, as well as mean ratings for each combination. Note that all same gender comparisons were male-male comparisons. It can be seen that similarity ratings are highest for the comparatively few within-cluster and within-gender comparisons, second highest for across-cluster and within-gender comparisons, third highest for within-cluster and across-gender combinations, and lowest for across-cluster and across-gender comparisons. On first sight, there appears to be a fairly large effect of gender and a more moderate effect of cluster membership.

Table 1: *Distribution of ratings across clusters and genders*

Contour comparison	n	mean	SD
Different cluster, different gender	964	1.86	0.95
Different cluster, same gender	756	3.04	1.03
Same cluster, different gender	293	1.94	0.93
Same cluster, same gender	84	4.79	0.49

Fig. 3 shows raw counts of ratings broken down by across- and within-cluster comparisons. It can be seen that “identical” ratings were more common for the within-cluster comparisons on the right. Since we did not originally plan to test any within-cluster comparisons at all, we will focus on exploratory analyses in the remainder of this paper.

However, we wanted to test whether there were *only* gender effects, or whether there was a main effect of cluster membership as well. To that end, we fitted a Bayesian cumulative link mixed model using R package `brms` [21], with cluster combination, gender combination and their interaction as predictors of similarity ratings (random intercepts for participants and items, and random by-participant slopes for cluster combination and gender combination). The results indicate a robust effect of cluster membership, such that within-cluster comparisons received higher similarity ratings than across-cluster comparisons ($\delta = 1.7$, 95% CI $[1.31, 2.1]$, $P(\delta > 0) = 1$).

Fig. 4 presents the 12 contour pairs with the highest mean similarity ratings, ordered from most similar in the top left to less similar in the bottom right. Note that the four most similar contour pairs were within-cluster comparisons – as a matter of fact, these were even within-speaker comparisons. The very high ratings for these contour pairs indicate that, while the task may have been difficult, participants were able to detect differences and similarities between contour pairs. Fig. 4 also reveals another two aspects of the results: As may be gleaned from the y-axis, all of the contour pairs pictured here are within-gender, male-male comparisons. Furthermore, only the within-speaker comparisons received very high ratings – the 12th most similar contour pair already received a comparatively low mean similarity rating of 3.3 (although bear in mind that the mid-point of the scale was labeled “similar”, so these pairs were probably still perceived as rather similar).

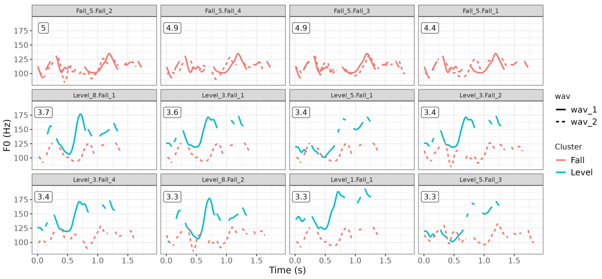

Figure 4: *The 12 most similar contour pairs, sorted by mean similarity rating. Mean similarity ratings are shown in boxes.*

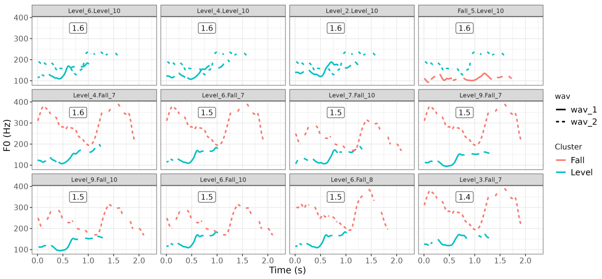

Figure 5: *The 12 least similar contour pairs, sorted by mean similarity rating. Mean similarity ratings are shown in boxes.*

Fig. 5 presents the 12 *least* similar contour pairs. First, note that all of these contour pairs are across-gender comparisons. The three within-cluster comparisons in the top left have a fairly similar pitch register and slight differences in duration, while the eight comparisons involving three different late-falling contours feature fairly extreme differences in pitch register and also quite large differences in duration. There are also differences in pitch movement in the first half of the utterance, such that the late-falling contours are also more steeply falling than the Level contours early on. So while it is clear that these late-falling contours stuck out in perception, there is a confound of several different variables. Ultimately, follow-up work will have to disentangle the factors pitch register, (location and steepness of) pitch movement, and duration.

Fig. 6 shows the negative correlation between the absolute difference (in semitones) in mean F0 between members of a contour pair and the mean similarity rating. Two main results are visible here. First, for the (originally intended) across-cluster comparisons shown as green symbols, there is a large effect of gender and/or F0 difference: Contour pairs were rated as more similar when the contour pairs were both produced by male speakers and when the absolute difference in mean F0 was below 6 ST than when the contour pairs were produced by a

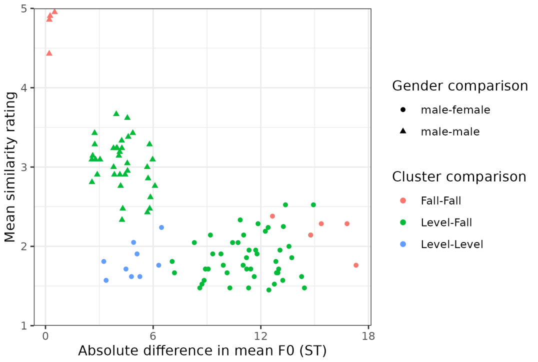

Figure 6: *Correlation between absolute difference in mean F0 in semitones and mean similarity ratings*

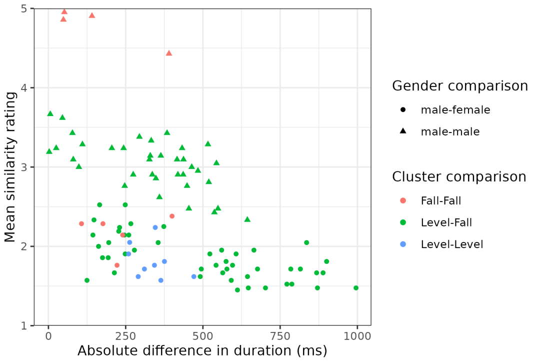

Figure 7: *Correlation between absolute difference in duration in milliseconds and mean similarity ratings*

male and a female speaker and F0 difference was above 6 ST. A similar pattern emerges for the Fall-Fall comparisons shown as red symbols, with the caveat that the within-gender comparisons were also all within-speaker. Second, the Level-Level comparisons, despite mostly featuring F0 differences below 6 ST, received low overall similarity ratings. Crucially, all of these comparisons were across-gender, so conclusions about the perceptual validity of the Level cluster are hard to draw. That said, note that the across-gender Fall-Fall comparisons (red circles) received slightly higher similarity ratings than the across-gender Level-Level comparisons (blue circles), despite the former comparisons featuring much more extreme differences in F0. This suggests that F0 register is not the sole determinant of perceived similarity. At any rate, the results indicate that participants were unable or unwilling to follow the instructions to disregard differences between speakers and/or genders.

Another factor that influenced perceived similarity was duration, or rather duration differences between the two members of a contour pair. Fig. 7 shows mean ratings as a function of absolute differences in duration. Besides the overall negative correlation between absolute differences in duration and mean similarity ratings, the effect is most clearly visible for the within-speaker comparison (red triangles) with the largest duration difference: It received noticeably lower similarity ratings than the other three within-speaker comparisons with more similar durations.

4. Discussion & conclusion

Overall, we take the present results to indicate that the (fairly abstract) task of judging the perceived similarity of hummed contours can be performed by phonetically untrained listeners of German, and that the original cluster analysis yielded valid results with respect to these two clusters. We nevertheless want to be careful in drawing phonological conclusions. While the putative H-L% boundary tones did stick out in perception and received low similarity ratings when compared to less steeply falling plateaus, they differed from other contours along, essentially, *all* the prosodic dimensions left in the stimuli: Pitch movement, pitch register and duration. From the present study, we can certainly say that this *bundle* of dimensions is involved in creating perceived dissimilarity for these contours. Future work will disentangle the relative contributions of each dimensions.

It should be pointed out that the original cluster analysis had no information about raw contour duration or speaker-specific F0 register: All contours were normalized to the same length, and F0 was normalized within-speaker to semitones relative to each speaker's median pitch. It is therefore perhaps not surprising that we found an influence of duration and speaker F0 register on perceived similarity that, to a certain extent, cuts across cluster membership. What *was* unexpected was the influence that speaker gender seemingly had independent of F0 register: Across-gender, within-cluster pairs received low similarity ratings even if the difference in F0 register was comparatively small. The question of whether this was actually a direct effect of speaker gender, or whether there were other, confounding factors, must be left for future work. We intend to reduce the number of comparisons to 5*5, while more systematically comparing contours across and within clusters, so that every contour is compared to every other contour, including itself. This will allow more robust conclusions about the relative contributions of cluster membership, F0 register, gender, and duration.

With respect to conclusions about the phonological questions raised in the introduction, regarding the issues of continuum vs. categories and the mapping of form to meaning in intonation, we want to exercise caution: In this study, we have only shown that German listeners are able to distinguish F0 contours presented in highly abstracted form, and that, in performing this task, they are sensitive to both (presumed) categorical distinctions, such as cluster membership, and continuous distinctions, such as duration and F0 register. Drawing conclusions about the form-meaning relationship within these particular contours, however, would require direct judgments about their semantic-pragmatic functions. It has been shown for concrete realizations of plateau contours in German excuses that something very close to categorical perception of pragmatic function exists within the span of just a few semitones [22]. Whether a similar task could be done with abstracted contours is a question that we must leave open.

5. Acknowledgements

The research for this paper has been funded by the German Research Foundation (DFG) – Project-ID 281511265 – as part of the SFB 1252 “Prominence in Language” in the projects A03 and A06 at the University of Cologne. The authors wish to thank Aleksandra Kopotilova for her contributions to the perception study.

6. References

[1] J. Pierrehumbert and J. Hirschberg, “The meaning of intonational contours in the interpretation of discourse,” in *Intentions in Communication*, P. R. Cohen, J. L. Morgan, and M. E. Pollack, Eds. Cambridge, MA: MIT Press, 1990, pp. 271–311.

[2] B. Braun, “Phonetics and phonology of thematic contrast in German,” *Language and Speech*, vol. 49, pp. 451–493, 2006.

[3] K. J. Kohler, “Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions,” in *From Traditional Phonology to Modern Speech Processing*, G. Fant, H. Fujisaki, J. Cao, and Y. Xu, Eds. Beijing: Foreign Language Teaching and Research Press, 2004, pp. 205–215. [Online]. Available: https://www.ipds.uni-kiel.de/kjk/pub_exx/aipuk35a/aipuk35a_5.pdf

[4] M. Grice, S. Ritter, H. Niemann, and T. B. Roettger, “Integrating the discreteness and continuity of intonational categories,” *Journal of Phonetics*, vol. 64, pp. 90–107, 2017.

[5] G. Demenko and A. Wagner, “The stylization of intonation contours,” in *Proceedings of Speech Prosody 2006*, Dresden, Germany, 2006, paper 254.

[6] J. Cole, J. Steffman, S. Shattuck-Hufnagel, and S. Tilsen, “Hierarchical distinctions in the production and perception of nuclear tunes in American English,” *Laboratory Phonology*, vol. 14, no. 1, pp. 1–51, 2023.

[7] T. J. Laméris, K. K. Li, and B. Post, “Phonetic and phonolexical accuracy of non-native tone production by English-L1 and Mandarin-L1 speakers,” *Language and Speech*, vol. 66, no. 4, pp. 974–1006, 2023.

[8] F. Jabeen and P. Wagner, “Variability in Punjabi semi-spontaneous narrative speech: An automatic clustering based analysis,” in *Proceedings of Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, S. Betz, B. Ludusan, and P. Wagner, Eds., Bielefeld, 2023, pp. 71–75.

[9] K. K. Li, F. Nolan, and B. Post, “Clustering lexical tones with intonation variation,” in *Proceedings of the Second Interonational Conference on Tone and Intonation*, M. Dong, Y. Lu, and R. Jian, Eds. Singapore: COLIPS, 2023, pp. 87–88. [Online]. Available: https://www.colips.org/conferences/tai2023/proceedings/pdf/2023.tai-abstract.43.pdf

[10] C. Kaland, “Intonation contour similarity: f0 representations and distance measures compared to human perception in two languages,” *Journal of the Acoustic Society of America*, vol. 154, no. 1, pp. 95–107, 2023.

[11] C. De Looze and D. Hirst, “The OMe (Octave-Median) scale: a natural scale for speech melody,” in *Proceedings of Speech Prosody 7*, N. Campbell, D. Gibbon, and D. Hirst, Eds., Dublin, 2014, pp. 910–914.

[12] H. Seeliger and C. Kaland, “Boundary tones in German wh-questions and wh-exclamatives – a cluster-based approach,” in *Proceedings of Speech Prosody 2022*, S. Frota, M. Cruz, and M. Viagário, Eds., Lisbon, 2022, pp. 27–31.

[13] S. Repp and H. Seeliger, “Contrast + givenness, local + non-local. The influence of complex information-structural settings on the prenuclear, nuclear and post-nuclear regions in exclamatives and questions,” *Language and Speech*, subm.

[14] C. Kaland, “Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours,” *Journal of the International Phonetic Association*, pp. 1–30, 2021.

[15] C. Kaland and T. M. Ellison, “Evaluating cluster analysis on f0 contours: An information theoretic approach on three languages,” in *Proceedings of the 20th ICPhS*, R. Skarnitzl and J. Volín, Eds. Prague: Guarant International, 2023, pp. 3448–3452. [Online]. Available: https://guarant.cz/icphs2023/108.pdf

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[17] M. Grice, S. Baumann, and R. Benzmüller, “German intonation in autosegmental-metrical phonology,” in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford: Oxford University Press, 2005, pp. 55–83.

[18] D. R. Ladd, “Stylized intonation,” *Language*, vol. 54, no. 3, pp. 517–540, 1978.

[19] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” 2021. [Online]. Available: http://www.praat.org

[20] S. Mathôt, D. Schreij, and J. Theeuwes, “OpenSesame: An open-source, graphical experiment builder for the social sciences,” *Behavior Research Methods*, vol. 44, no. 2, pp. 314–324, 2012.

[21] P.-C. Bürkner, “brms: An R package for Bayesian multilevel models using Stan,” *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[22] O. Niebuhr, “Resistance is futile – the intonation between continuation rise and calling contour in German,” in *Proceedings of INTERSPEECH 2013*, F. Bimbot, C. Fougeron, and F. Pellegrino, Eds., Lyon, France, 2013, pp. 225–229.