



# Perception compared to clustered intonation variation in German *wh*-questions

Heiko Seeliger<sup>1</sup>, Constantijn Kaland<sup>2,3</sup>

<sup>1</sup>Department of German Language and Literature I, University of Cologne, Germany

<sup>2</sup>Ifl Phonetik, University of Cologne, Germany

<sup>3</sup>Linguistics I – Phonetics and Phonology, University of Düsseldorf, Germany

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

## Abstract

This article presents the results of a perception study using hummed F0 contours as stimuli. The original contours were produced in a production study investigating German *wh*-questions and *wh*-exclamatives, and had previously been classified using cluster analysis. A previous perception study on the same dataset had found that the output of the cluster analysis was predictive of perceived contour similarity, but that contour duration and F0 register (both of which were normalized for the original cluster analysis) also had a large influence.

In this study, we normalized the stimuli in terms of duration and F0 register. Original cluster membership was not predictive of perceived similarity, suggesting that the original clusters did not form ‘pure’ perceptual classes when correcting for duration and register. However, cluster analysis applied directly to the subset of stimuli did result in a clustering that was predictive of perceived similarity. We also investigated correlations between acoustic differences and perceived similarity for the four quarters of each contour pair. Results show that only later acoustic differences have an impact on perception.

**Index Terms:** intonation, F0, pitch, cluster analysis, perception

## 1. Introduction

Cluster analysis (CA) of F0 contours is an emerging method in the analysis of tone and intonation. While CA has been applied in research of prosody for decades (e.g., [1]), this usually took the form of it being applied on aggregated data, e.g., on per-syllable means, or on perception ratings. Lately, CA has also been used to analyze ‘raw’ F0 contours directly (see [2, 3, 4] for background). Recent studies have investigated a wide range of phenomena using this method: post-lexical intonation in Cantonese [5], the Mainstream American English ToBI inventory [6], the impacts of sarcasm [7] and emotion [8] on intonation in American English, boundary tones [9] and backchannels [10] in German, filler particles across languages [11], focus marking in Mexican Spanish ethnolects [12], and intonation and tone in an under-researched Papuan language [13].

Cluster analysis necessarily analyzes F0 contours, i.e., subsequent values of (the inverses of) glottal periods. Ultimately, however, we are interested in *pitch* contours, i.e., in the psycho-acoustic percept of F0 (cf. [14] for discussion of this distinction). This ultimate aim requires an investigation of the link between F0 and pitch, or between production and perception. A recent study [3] investigated correlations between various F0 measures, distance measures and two listener groups (German and Papuan Malay), and showed that contour perception was quite similar between the two speaker groups. However, correlations between the acoustic difference measures and perceived similarity were only moderate, suggesting that there are still unknown factors

influencing F0 perception, e.g., non-F0 cues or phrasal context.

In this study, we build on and refine the methodology of a previous perception study [15] using clustered F0 contours [9] as stimuli. We show that cluster analysis can yield results that are in line with the outcome of a perception study, or in other words: cluster analysis can yield perceptually distinct clusters.

## 2. Background

[9, 15] investigated intonational variation in an experimental corpus of German *wh*-questions and *wh*-exclamatives (from [16]; example lexicalization: *wo die schon überall Germanen erforscht hat*, ‘where she has already researched Germanic peoples’). [9] focused on the question of whether CA can result in phonologically meaningful clusters when applied to a relatively large dataset of German read speech (number of utterances: 1109). A special focus lay on the boundary tones, i.e., F0 patterns related to the edges of phonological phrases. Two separate cluster analyses were combined in that study: one CA on the entire utterances and one CA on the final two syllables of all utterances (which gave ‘extra weight’ to utterance-final pitch). The combined analysis performed well in separating clusters that mostly corresponded well to ‘canonical’ GToBI [17] nuclear contours, such as L\* H-^H% or L+H\* L-H%.

Two clusters were singled out for further analysis in [9], as they did *not* appear to straightforwardly correspond to one canonical GToBI nuclear contour: one cluster containing many late, seemingly boundary-related falls and one cluster containing medium-high plateaus, some of them downstepped. Note that the two clusters appeared from the analysis of the final two syllables, not from the analysis on the whole contour. [9] proposed that the falling cluster consisted of instances of H-L%, which is not officially a part of GToBI (and should not be confused with the boundary tone of the same name in MAE-ToBI, which instead corresponds to GToBI H-%), while the level cluster seemed to fall in between H-% and !H-%. !H-% is the boundary tone occurring in the German calling contour (cf. [18, 19]), but note that none of the instances of it in this dataset intuitively sounded like calling contours. In the remainder of this article, we will refer to the falling cluster as F and to the level cluster as L for short, so as not to prejudice a phonological analysis.

[15] ran a perception study using a selection of contours from the F and L clusters as stimuli. Only utterances that were originally produced as *wh*-questions were selected. Contours were played in pairs and then rated by participants on a 5-point similarity scale. The contours were presented as hummed tones, i.e., they contained no lexical or segmental information, consisting entirely of a schwa-like vowel. Besides this removal of lexical and segmental information, no changes were made to the F0 contours, i.e., differences in duration and F0 register were

not corrected for. The study found that within-cluster comparisons were indeed judged to be more perceptually similar than between-cluster comparisons, but there were several confounds. First, the most similar pairs were not just within-cluster comparisons, but also within-speaker (and hence also within-sex) comparisons. Second, there was a large impact on perceived similarity from differences in duration and F0 register (both of which were not ‘visible’ to the original cluster analysis, since contours are scaled to the same length for cluster analysis and the contours in [9] had their F0 normalized within-speakers). Most crucially, the contours in the F cluster differed from the more level contours of the L cluster in all three aspects, i.e., F0 movements, F0 register / speaker sex, and duration.

Thus, while it seemed likely that the original cluster analysis indeed resulted in perceptually distinct clusters, we could not isolate the relative contributions of F0 shape, F0 register and contour duration. These confounds motivated the present study. We set out to make exhaustive comparisons between contours that have been normalized with respect to F0 register and their total duration, so that only differences in the locations and magnitudes of F0 movements remain as cues. Crucially, these are the same cues that are available to a cluster analysis of F0 contours.

### 3. Methodology

#### 3.1. Stimulus selection and normalization

We made the following changes to the methodology relative to [15]: (i) selection of 5 contours from each cluster, for a total of 10 single contours (down from 10 contours from each cluster); (ii) selection of contours that were balanced for original speaker and gender as far as possible (instead of selecting the most typical contours from each cluster); (iii) where possible, inclusion of identical speakers in both clusters; (iv) presentation of all possible combinations of contours, including presentation of contour pairs consisting of identical contours (instead of mostly between-cluster comparisons); (v) normalization of duration of each contour to the mean duration of all 10 contours; (vi) normalization of F0 register to the median F0 value of all 10 contours. The main goal of F0 register normalization was to make speaker gender unavailable as a perceptual cue.

Table 1 gives an overview of information about the five contours that we picked from each cluster. Regarding original nuclear accent locations, 9 of the utterances featured nuclear accents on the object, which was roughly located starting at the halfway point of each utterance (the stressed syllable was followed by four unaccented syllables). One utterance featured a nuclear accent on the lexical verb, whose stressed syllable was the penultimate syllable of the utterance. We come back to this potential ‘odd one out’ in the Discussion. Finally, again only original questions were selected.

	Fall	Level
Speakers	f2 (x2), <u>f3</u> , <u>m3</u> , m4	f1, <u>f3</u> , m1, m2, <u>m3</u>
Nuclear accent	Obj. (x4), Verb (x1)	Obj. (x5)
Duration (orig.)	1.76 s	1.69 s
Duration (norm.)	1.72 s	1.72 s
F0 median (orig.)	203.21 Hz	160.83 Hz
F0 median (norm.)	176.02 Hz	176.02 Hz

Table 1: Stimulus metadata by cluster. ‘f’ denotes female speakers, ‘m’ male speakers. Underlined speakers are present in both clusters.

The F0 normalization was carried out using a custom Praat [20] script, using to the following equation:

$$Norm(F0_i) = \frac{F0_i \cdot \text{median}(F0_{\text{data}})}{\text{median}(F0_{\text{contour}})} \quad (1)$$

Calculation by means of multiplication and division ensures correct scaling of the F0 range: contours that are normalized into a lower register must also be ‘squeezed’ vertically in order to be perceptually equivalent to the original contour, and vice versa for normalization into a higher register and ‘stretching’ (see Fig. 1 for illustration).

The order of operations was as follows: (1) normalization of the duration of the original sound using PSOLA, (2) interpolation of the pitch contour (i.e., filling in unvoiced segments), (3) smoothing of the pitch contour (10 Hz bandwidth), (4) calculation of the median F0 of the resulting contour, (5) transposition of the contour to the normalized F0 register, using the formula given in Eq. 1, (6) creation of a new sound file from the resulting Pitch object using the Praat function `To Sound (hum)`. The Praat script and all stimuli are available at <https://osf.io/fe6w4/>.

Acoustic measures related to duration and F0 register before and after normalization are given in Table 1. Note that cluster F was normalized downwards, originally containing 3 female and 2 male speakers, and vice versa for cluster L. After normalization, our auditory impression of the stimuli was that they were genuinely ambiguous with respect to original speaker gender.

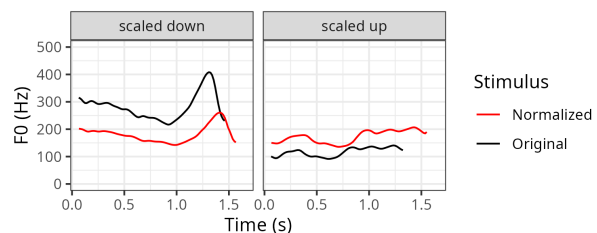


Figure 1: Two example contours before (black line) and after (red line) normalization of duration and F0 register. The contour on the left is an F contour by a female speaker; the one on the right an L contour by a male speaker.

#### 3.2. Procedure

The experiment was run using PsyToolkit [21, 22]. 27 native listeners of German participated, who were recruited from the student population of the University of Cologne. Contours were played in pairs, and listeners then judged the perceived similarity of each pair on a five-point scale (the end points were labeled *identisch*, ‘identical’, and *unähnlich*, ‘dissimilar’). The instruction made reference to the *Satzmelodien*, ‘sentence melodies’, so intonation was explicitly highlighted (in layman friendly terms), but any potential functions of intonation, such as communicative intentions, were not. Each listener judged all 55 contour pairs. There were no fillers. The order of stimulus presentation was randomized for each participant. In total, the experiment resulted in 1485 similarity ratings. We excluded comparisons of identical contours from the statistical analysis, which left 1215 ratings. For the statistical analysis, we fitted cumulative link mixed models (CLMMs) using R package `ordinal` [23]. Both of the models we report feature by-subject and by-stimulus

random intercepts and by-subject random slopes for the predictor, and they both used treatment contrasts with between-cluster comparisons as the intercept.

## 4. Results

### 4.1. Direct comparison of original clusters

The results of the perception study are shown in Fig. 2. The visually most striking result is that participants were clearly able to identify pairs of identical contours, indicated by the vast majority of ‘identical’ ratings (5) on the right side of the plot. Turning to the comparisons of different contours, original cluster membership did not have a significant impact on perceived similarity. While there is a slight trend for F-F and L-L comparisons to receive more ‘identical’ ratings than the between-cluster F-L comparisons, this was not significant (all  $p > 0.1$ ).

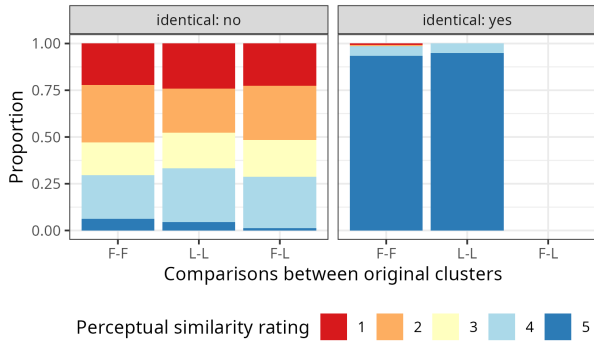


Figure 2: Distribution of similarity ratings by membership in original clusters. Pairs of different stimuli are shown on the left, pairs of identical stimuli on the right.

### 4.2. Post-hoc analyses

Since the original cluster analysis clustered all 1109 contours from [16], and since the two clusters investigated here represented ‘mixed’ clusters (i.e., they fell into the same cluster based on the whole contour, but different clusters based on the final two syllables), we wanted to investigate to which extent cluster analysis applied *only* to this subset of 10 contours would correlate with perceived similarity. To this end, we ran a cluster analysis on the 10 stimuli (using F0 in Hz – since the stimuli were already normalized – with a euclidean distance metric and hierarchical agglomerative clustering with complete linkage).

The best clustering solution consisted of two clusters: one cluster consisting of the three late-falling contours from the F cluster and one cluster consisting of the seven other stimuli, see Fig. 3. An analysis using three clusters still clustered the three late-falling contours into the same cluster, so for ease of presentation we only discuss the cluster analysis using two clusters. We refer to the original cluster analysis (1109 contours) as ‘original’ and to the results of the cluster analysis on the 10 stimuli as ‘subset clusters’, using the numeric labels shown in Fig. 3 to refer to the subset clusters.

Fig. 4 shows the similarity ratings split by subset cluster membership. It can be seen that the three late-falling contours (labeled ‘2’ in Figs. 3 and 4) were judged to be more similar to each other than comparisons between subset clusters (labeled ‘1-2’). For the more level contours in subset cluster ‘1’, a similar picture obtains, although the within-cluster comparisons are not

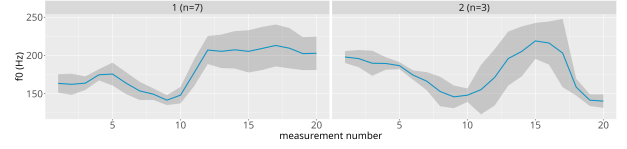


Figure 3: Results of the cluster analysis of the 10 stimuli

as similar as for the late-falling contours. A CLMM predicting perceived similarity from subset cluster membership confirms these effects (1-1 vs. 1-2:  $p < 0.05$ ; 2-2 vs. 1-2:  $p < 0.01$ ).

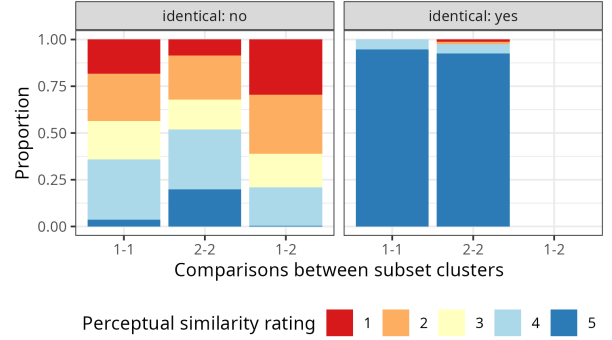


Figure 4: Distribution of similarity ratings by membership in the subset cluster analysis. Pairs of different stimuli are shown on the left, pairs of identical stimuli on the right. Cluster labels refer to the labels shown in Fig. 3.

### 4.3. Locus of perceived differences

We also investigated *which* differences between contours had larger impacts on perceived similarity. To this end, we calculated two measures of acoustic similarity for the four quarters of each contour pair separately. The two measures were F0 range and  $\Delta F0$ . F0 range was a simple measure of the difference between maximum F0 and minimum F0 within each quarter of *one* contour; Fig. 5 shows the absolute differences between contour pairs in F0 range on the x-axis and mean perceived similarity on the y-axis.

$\Delta F0$  was calculated as follows:

$$\Delta F0 = \frac{1}{n} \sum_{i=1}^n |F0_i - F0_{i+1}| \quad (2)$$

This is a measure of the mean difference between adjacent F0 points (in Hz). Fig. 6 shows the absolute differences between members of a contour pair in terms of  $\Delta F0$  on the x-axis and mean perceived similarity on the y-axis.

For both F0 range and  $\Delta F0$ , it can be seen that later differences have a larger impact on perceived similarity. Only in the third and fourth quarter are there correlations between acoustic similarity and perceived similarity. Early differences in the first half of the contour pairs do not seem to have much of any impact on perceived similarity. Note that early differences do exist and are of comparable magnitude to late differences, so the greater impact of later acoustic difference is not simply an epiphenomenon of later acoustic differences being larger.

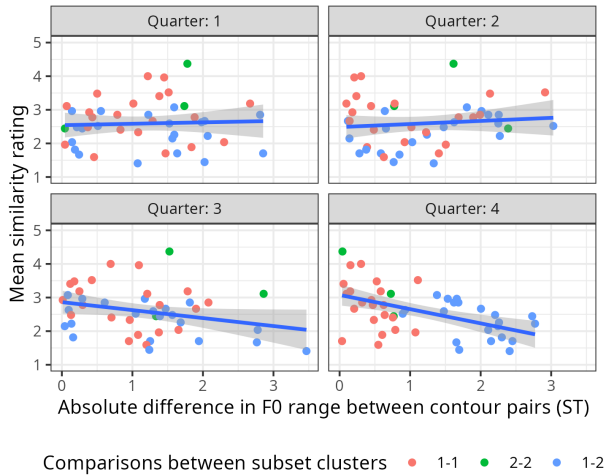


Figure 5: Correlations between  $F_0$  range (x-axis) and mean perceived similarity (y-axis), shown separately for the four quarters of each contour pair. Every dot is one contour pair. Identical pairs have been excluded from this plot.

## 5. Discussion

Our results are two-fold: first, the original clusters from [9] were indeed ‘mixed’. Given that these clusters were the results of a combination of two separate cluster analyses on different time domains, this finding is maybe not surprising. Second, despite this mixture being present in the stimuli, cluster analysis performed directly on the 10 stimuli themselves is successful in forming perceptually distinct clusters, as shown by the results of the subset cluster analysis.

We nevertheless want to advise caution, in particular with respect to drawing wide-reaching phonological conclusions about German intonational phonology. The reason for this is that the stimulus preparation necessarily stripped the stimuli of all lexical or segmental cues to stress, accentuation and alignment (in terms of tonal targets and syllables). To give a concrete example of what this entails, consider the late-falling contours labeled ‘2’ in Fig. 3: we cannot say whether these contours were perceived to be more similar to each other because the listeners all tended to perceive, say, an H-L% boundary tone (following a low-rising accent), or because the listeners all tended to perceive a late and perhaps an unusually wide peak accent. With pure  $F_0$  contours that contain no clear cues to syllable structure, these differences in alignment will be hard to impossible to perceive. As a matter of fact, the contour pair with the highest mean similarity rating (among non-identical contour pairs) is a pair involving the one stimulus with a late nuclear accent on the verb, which was compared to a ‘boxy’ plateau following a low accent on the object. That this pair received the highest mean similarity rating suggests that accentual peaks and boundary-related (falling) plateaus were perceived similarly.

Regarding the finding that listeners are more sensitive to acoustic differences later in the utterance, we remain agnostic on whether this sensitivity represents a pure recency effect, or might be taken to represent a sensitivity specifically to the nuclear part of the contours. We consider a recency effect to be more likely, since – as we just discussed – it is not clear that the stimuli contained any cues to what constituted the nuclear part of each contour, besides the perhaps trivial cue that nuclear

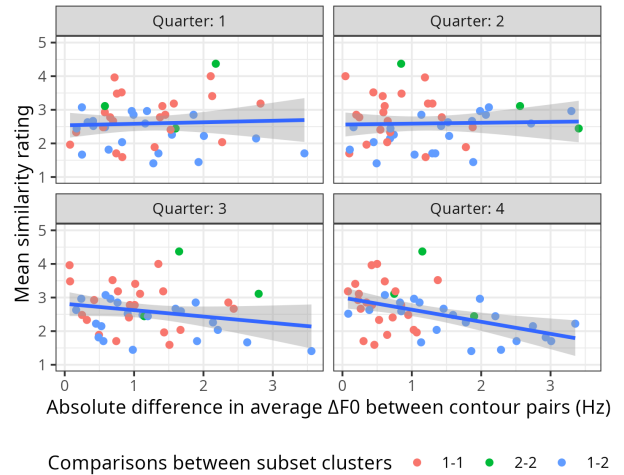


Figure 6: Correlations between  $\Delta F_0$  (x-axis) and mean perceived similarity (y-axis). All other aspects are like in Fig. 5.

accents tend to occur late in utterances by definition (nuclear accents being the final accent in a phrase). Even then, a late  $F_0$  movement could instantiate an accent or a boundary tone, as just pointed out, and these two phonological objects might have different perceptual effects, at least in the presence of lexical and segmental information (cf. [24]). That said, it is also possible that, in general, the nuclear part of an intonational contour is important to perception *because of* recency effects, which would mean that the two concepts cannot be disentangled.

## 6. Conclusion

With the caveats we just presented in mind, the present study shows (a) that phonologically untrained listeners of German can successfully perform the relatively abstract task of giving ratings of perceived similarity for hummed contour pairs, and (b) that cluster analysis ‘agrees with’ these listeners to a substantial extent. Taken together, these findings suggest that  $F_0$  contours *can* be analyzed holistically, since CA and listeners arrive at comparable categorizations of contours. Although this categorization is not representative of natural language nor of natural intonation, it may very well be a task that listeners do continuously (alongside integrating all cues that we stripped off).

The findings of the present study raise the issue of *what* to cluster – only CA applied directly to the set of 10 stimuli is in line with the perception of participants, while the original CA applied to the entire dataset was not. The results indicate that focusing closer on the phenomenon of interest yields better results. In other words, the original cluster analysis, in clustering the entire dataset, lost sensitivity for differences in this subset of 10 contours. That said, CA can still be helpful to semi-automatically sort a large dataset.

We close by pointing out two potential avenues for future research: the potential recency effect could be tested by assessing listener’s working memory capacity (WMC). If listeners with weaker WMC show stronger correlations between late differences and perceived similarity, this would indeed point towards a recency effect. Regarding the stimuli, it might be possible to use the amplitude envelope in order to construct stimuli that keep the syllabic structure (more) intact and potentially retain cues to stress, while still being free of lexical information.

## 7. 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. CK was funded by the German Research Foundation (DFG) – Project-ID 281511265 and 559664412. The authors thank the audience at TAI 2025 for comments and suggestions.

## 8. References

- [1] J. 't Hart, R. Collier, and A. Cohen, *A perceptual study of intonation*. Cambridge: Cambridge University Press, 1990.
- [2] 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.
- [3] —, “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.
- [4] C. Kaland, J. Steffman, and J. Cole, “K-means and hierarchical clustering of f0 contours,” in *Proceedings of Interspeech 2024*. Kos: ISCA, 2024, pp. 1520–1524.
- [5] K. K. Li, F. Nolan, and B. Post, “Clustering lexical tones with intonation variation,” in *Proceedings of the Second International 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>
- [6] J. Steffman, J. Cole, and S. Shattuck-Hufnagel, “Intonational categories and continua in American English rising nuclear tunes,” *Journal of Phonetics*, vol. 104, p. 101310, 2024.
- [7] C. Tatár, J. Brennan, J. Krivokapić, and E. Keshet, “Examining melodiousness in sarcasm: wiggleness, spaciousness, and contour clustering,” in *Speech Prosody 2024*. Leiden: ISCA, 2024, pp. 677–681.
- [8] D. R. Turner, “Intonation through emotion: Evidence of form and function in American English,” Ph.D. dissertation, Northwestern University, Evanston, 2025.
- [9] 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. Vigário, Eds. Lisbon: University of Lisbon, 2022, pp. 27–31.
- [10] E. Möking, S. Wehrle, C. Kaland, and M. Grice, “Intonational dynamics of German backchannels across conversational contexts,” in *Phonetics and Phonology in Europe 2025*, Palma de Mallorca, 2025.
- [11] M. Böttcher and C. Kaland, “Prosodic prototypes of filler particles across three languages,” in *12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)*. Lisbon: ISCA, 2025, pp. 12–16.
- [12] G. Marchini and J. Steffman, “Data-driven approaches to pitch modelling in two Mexican Spanish ethnolects: K-means clustering & GAMMs,” in *Proceedings of Interspeech 2025*. Rotterdam: ISCA, 2025, pp. 2940–2944.
- [13] A. Zahrer, “Exploring natural speech intonation of an under-researched Papuan language,” in *Speech Prosody 2024*. Leiden: ISCA, 2024, pp. 1095–1099.
- [14] B. Ahn, N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos, “PoLaR annotation guidelines,” 2022.
- [15] H. Seeliger, A. Lützel, and C. Kaland, “The perception of German wh-phrase-final intonation,” in *Proceedings of the Second International Conference on Tone and Intonation (TAI)*, Singapore, 2023, pp. 10–14.
- [16] 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,” *Advance*, 2024.
- [17] M. Grice, S. Baumann, and R. Benz Müller, “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] 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.
- [20] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” 2025. [Online]. Available: <http://www.praat.org>
- [21] G. Stoet, “Psytoolkit: A software package for programming psychological experiments using linux,” *Behavior Research Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.
- [22] —, “Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments,” *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.
- [23] R. H. B. Christensen, *ordinal—Regression Models for Ordinal Data*, 2023, R package version 2023.12-4.1. [Online]. Available: <https://CRAN.R-project.org/package=ordinal>
- [24] M. Lialiou, M. Grice, C. T. Röhr, and P. B. Schumacher, “Auditory processing of intonational rises and falls in German: Rises are special in attention orienting,” *Journal of Cognitive Neuroscience*, vol. 36, no. 6, pp. 1099–1122, 2024.