INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

In-group advantage in the perception of emotions: Evidence from three varieties of German

Moritz Jakob[1], *Bettina Braun*[1] *and Katharina Zahner-Ritter*[2]

[1]University of Konstanz, Department of Linguistics, Germany
[2]University of Trier, Department II, Phonetics, Germany
{moritz.jakob,bettina.braun}@uni-konstanz.de, k.zahner-ritter@uni-trier.de

Abstract

Various studies on the perception of vocally expressed emotions have shown that recognition rates are higher if speaker and listener belong to the same cultural or linguistic group. This so-called *in-group advantage* is commonly attributed to prosodic differences in the expression of emotion across groups. Evidence comes mostly from using cross-linguistic and/or cross-cultural study designs. Previous research suggests that varieties of German differ in their use of prosody and can be discriminated based on prosodic features alone. In this paper, we tested whether emotion recognition rates differ across varieties of German: Listeners from three dialectal areas (Hamburg, Vienna, Zurich) identified emotions on semantically neutral sentences (choosing between anger, happiness, relief, surprise or "other"), spoken by actors from the three regions. Correctness rates show that emotions are recognized better if speakers and listeners are native speakers of the same variety. However, further analyses suggest that the in-group advantage does not surface consistently across individual emotions. To explain these results, the prosodic realization of the sentences was tested for interactions between emotion and variety. Here, intensity seemed to differ most across varieties and emotions. Importantly, we show that the in-group advantage extends from cultural groups to dialectal groups of a language.

Index Terms: emotion perception, in-group advantage, speech prosody, varieties of German

1. Introduction

Research on visual and vocal emotion perception suggests that people recognize emotions better if they are expressed by "members of the same national, ethnic, or regional group" [1, p. 203]. There is cross-linguistic and cross-cultural evidence supporting this claim [2–5]: For instance, Caucasian Americans had higher recognition rates than Native Americans for four emotions that were vocally expressed by Caucasian Americans. Conversely, the Native American perceivers outperformed Caucasian Americans in discriminating the same emotions if they were spoken by speakers of Cree [2]. [4] allude to a trend that recognition rates decrease with an increase in the distance of linguistic relationship (i.e., speakers of Germanic languages achieved higher recognition rates for emotions encoded in German than speakers of Romance languages, who in turn outperformed speakers of Bahasa Indonesian on the same task) [4].

The in-group advantage seems to result from a language-specific use of prosodic features to encode and decode emotions, e.g., intensity [5] and pitch [6] are used differently to encode certain emotions. Nevertheless, prosody marking emotions seems to be largely driven by universal principles: recognition rates are well above chance level for vocally expressed emotions even across different languages and cultures [2–5, 7].

While the in-group advantage has been extensively studied for different languages [3–5], little is known on whether it extends to different varieties of the same language (but see [7]). In the present study, we attempt to fill this gap by testing the perception of vocally expressed emotions in different varieties of German. Research on German dialects has revealed that varieties of German display differences in their general use of prosody: Pitch related cues, in particular, seem to be used distinctly by different varieties. Much of the exploration of this topic has focused on the timing of pitch accents [8–11]. However, cross-varietal differences in the F_0 range and speech tempo have also been found [12]. Furthermore, experiments conducted with recordings from which the segmental information had been removed revealed that listeners can perceptually distinguish varieties of German based on prosodic information alone [13, 14]. Referencing [15–18], [14] remark that the Swiss German varieties included in their perception study differed in the phonetic alignment of pitch accents, as well as in their articulation rate and vowel duration. [13] observe that their participants rarely confused speakers of Alemannic varieties with speakers from North-Western Germany. Similarly, speakers from the Austrian-Bavarian group were seldomly confused with speakers from Northern Germany. These patterns may be a related to the apparent North-South contrast in the realization of pitch accents that [8–11] point out in their analyses.

Given the differences in the use of prosody by speakers of different varieties of German, we set out to investigate whether the in-group advantage in vocal emotion perception prevails when the emotions are encoded by speakers of three varieties of German: Austrian German as spoken in Vienna (VN), Northern German as spoken in Hamburg (HH), and Swiss German as spoken in Zurich (ZH). A cross-varietal in-group advantage has so far been documented for speakers of varieties of English [7]. However, unlike the study described in [7], our study tests the perception of vocally expressed emotions by members of three groups that are culturally relatively closely related. In addition to the perception task, we also analyzed the recorded emotion expressions based on a set of eleven prosodic parameters related to intensity, pitch, speech tempo and voice quality. Our hypothesis is the following: Under the assumption that the in-group advantage transfers to varieties of a language, we expect that emotional expressions are recognized better if speakers and listeners are members of the same dialect group than if they are members of different dialect groups.

http://dx.doi.org/10.21437/Interspeech.2021-1172

2. Experiment

2.1. Methods

2.1.1. Participants

A total of 46 native speakers of German participated in a perception task. Seventeen of the participants grew up in or close to the city of Hamburg (M_{age}=28.1 years, SD=6.9 years, 10 females, 7 males), 17 grew up in the canton of Zurich (M_{age}=24.9 years, SD=3.1 years, 14 females, 3 males) and 12 grew up in or close to the city of Vienna (M_{age}=25.1 years, SD=4.5 years, 9 females, 3 males). Each participant rated their familiarity with each of the three different dialects on a scale from 1 (not at all familiar) to 5 (very familiar). The mean familiarity score was highest for the native variety in all three groups (HH: fam_{HH} = 4.2 > fam_{VN}= 2, fam_{ZH}=1.6; VN: fam_{VN}=4.8 > fam_{HH}= 2.3, fam_{ZH}=1.8; ZH: fam_{ZH}=5 > fam_{HH}=1.8, fam_{VN}=2.4).

2.1.2. Stimuli

Target sentences. To avoid an implicit emotional semantic bias of the sentences on which the emotions were encoded, we constructed (and pretested) twenty sentences for their emotional ambivalence. In the pretest, the sentences were presented in written form to 41 German native speakers (M_{age}=27.0 years, SD=6.1 years, 28 females, 13 males) using the online survey platform SoSci Survey [19]. Participants judged whether a sentence could be used to express eight emotions (anger, disgust, fear, happiness, pride, relief, sadness, and surprise). For each emotion, they chose between three options (*yes*, *no*, or *unsure*).

The three sentences with the highest overall *yes*-rates were used in the actual study. This resulted in the following three final test sentences: 'Es ist Amelie' (Engl. *It is Amelie*), 'Da kommt Gabriel' (Engl. *There comes Gabriel*), and 'Da steht Michael' (Engl. *There is [lit. stands] Michael*). The choice of which emotions were included in the experiment was also determined by the results of the pre-test. Four emotions were chosen that had a consistently high proportion of *yes*-answers for the three sentences (i.e., they had a *yes*-rate of >85% for each individual sentence): Anger, happiness, relief, and surprise.

Recordings. To compile the audio material for the study, 12 speakers were recorded expressing the three test sentences in the emotions anger, happiness, relief, and surprise. Four of the 12 speakers grew up in or close to the city of Hamburg (M_{age}=26.3 years, SD=6.2 years, 2 females, 2 males), four grew up in or close to the city of Vienna (M_{age}=23.5 years, SD=1.1 years, 2 females, 2 males), and four grew up in the canton of Zurich (M_{age}=32.5 years, SD=2.6 years, 2 females, 2 males). All speakers had attended drama school or were attending drama school at the time of recording. Most of them had previously worked in professional theatre or TV productions.

Recording sessions were supervised via videoconferencing tools. The speakers recorded themselves in a quiet environment using Praat [20] (44.1 kHz, 16-bit). Due to the remote recording conditions, equipment differed between speakers. The speakers were presented with a short description of a scenario for each sentence that triggered one of the four emotions. This was done to make the production more natural and to control the interpretation of the emotion by the encoders. Speakers were asked to silently read one scenario at a time and subsequently produce the sentence conveying a given emotion. The target emotion was explicitly stated before the test sentence (e.g., orig. 'Verärgert sagst du…', *angrily you say*…). The actors were asked to produce the sentence in their dialect but not to add any syllables (note that some dialects of German, e.g., Zurich German, naturally include a definite article before the name).

Data preprocessing. In total, there were 144 stimuli that were included in the perception task (4 emotions x 3 varieties x 4 speakers x 3 test sentences).

All recordings were judged to be of good quality by the first author. Given different recording hardware across speakers, the sound quality of the audio recordings and intensity differed slightly between participants. To account for the variation in intensity, the intensity of the audio files was normalized using Praat: In order to maintain intensity cues related to individual emotion categories, the recordings of the individual speakers were concatenated into a single file which was subsequently normalized to a mean intensity of 70 dB SPL and cut back into the separate sentences.

Procedure. In an alternative forced choice paradigm (AFC), participants rated test sentences (N = 48) for conveying different emotions.

The experiment was created and hosted online using the Gorilla Experiment Builder [21]. Participants were asked to perform the experiment on a laptop or desktop computer in a quiet environment with good speakers or headphones.

The AFC paradigm provided five alternatives: In addition to the four emotions that were included in the study (*anger*, *happiness*, *relief*, *surprise*) another response option was presented that allowed an open answer (*neither of the four but…*/orig. 'keine der vier sondern…'). Furthermore, participants were asked to rate how sure they were about their answer on a five-point Likert scale (1 = not at all sure, 5 = very sure). Each trial started with the presentation of a fixation cross in the center of the screen for 250ms, followed by the presentation of the test sentence in written form accompanied by the five answer alternatives of the AFC task, and the Likert scale. With 300ms delay, the audio stimulus was presented once (without a repeat option). Prior to the actual experiment, there was a short practice phase (N = 4 trials with four different emotions) to familiarize participants with the task. The stimuli presented in the practice phase included at least one example of each emotion, variety, and test sentence. None of the stimuli that were used in the practice phase were recorded by any of the 12 speakers whose recordings were used in the actual experiment.

Experimental conditions were distributed over items and speakers in a Latin square design. Accordingly, each participant only judged a subset of the 144 emotional expressions (N = 48). Three experimental lists were compiled in a pseudo-randomized order. The number of stimuli from each variety, speaker, emotion, and test sentence was identical in each list. No recording was included in more than one list meaning that each participant judged 48 emotional expressions, in total. The overall duration of the experiment was approximately 12-15 minutes.

2.2. Results

2.2.1. Perception Task

To have valid stimuli, we only evaluated responses to recordings that in-group members recognized with an accuracy of 40% or higher and hence well above chance. Responses to 96 recordings (36 HH, 27 VN, 33 ZH) were included in the dataset that was used in the analysis.

Table 1 summarizes the recognition rates that the three decoder groups (listeners) achieved for each of the four emotions spoken by members of the three varieties of German, as well as the total recognition rates.

Table 1: *Recognition rates of the three decoder groups for stimuli of each encoder group in percent. Bold values mark in-group judgements.*

Group	Anger	Happy	Relief	Surpise	Total
HH_{enc}					
HH_{dec}	**60.7**	**66.0**	**74.4**	**70.6**	**67.5**
VN_{dec}	70.0	63.0	62.5	72.2	67.4
ZH_{dec}	49.1	51.9	61.4	76.5	59.4
VN_{enc}					
HH_{dec}	44.2	41.9	54.5	82.0	58.2
VN_{dec}	**62.5**	**66.7**	**50.0**	**77.8**	**66.7**
ZH_{dec}	53.3	68.8	42.9	85.2	66.4
ZH_{enc}					
HH_{dec}	85.7	61.9	58.7	73.5	69.6
VN_{dec}	87.5	68.8	25.0	60.4	59.8
ZH_{dec}	**80.6**	**72.9**	**61.5**	**75.0**	**72.6**

Table 1 indicates that the overall accuracy rates for each encoder group were highest for decoders from the same varietal group. However, for the emotion expressions encoded by speakers from Hamburg and Vienna, the difference to the second highest accuracy rate was only small. The overall accuracy scores by speaker and listener condition are visualized in Figure 1. Considering the self-assessed familiarity scores (see 2.1.1) and the data presented in Table 1 and Figure 1, a higher familiarity with a non-native variety did not lead to a better recognition. Furthermore, none of the perceiver groups consistently achieved the highest accuracy rates for stimuli from their own variety across all four emotions. However, decoders from Zurich outperformed the other two groups on recognizing three out of the four emotions correctly if they were encoded by speakers from Zurich. Interestingly, perceivers from Vienna achieved the highest accuracy scores for *angry* stimuli across all three speaker groups. Similarly, perceivers from Zurich showed the highest recognition rates for *surprised* stimuli across all three speaker groups.

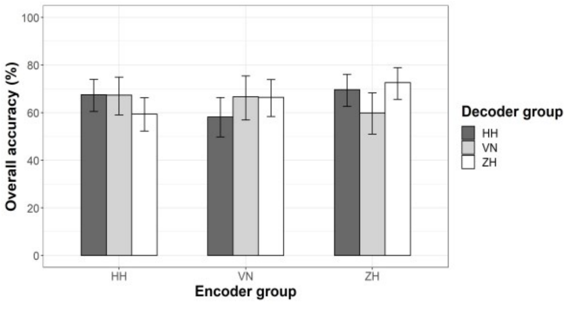

Figure 1: *Overall accuracy rates by encoder and decoder group conditions. Error-bars represent the 95%-CI.*

To statistically evaluate the in-group advantage, a two-level factor was created that separated the data in observations with *matching* group affiliation or *mismatching* group affiliation, cf. Figure 2. For instance, for the HH encoder group, ratings from the HH decoder group were coded as (*match*) "yes", ratings from the VN and ZH decoder groups as (*match*) "no".

Using the *lme4* R-package [22, 23], we fitted a generalized linear mixed model (GLMM) to analyze the effect of *match* on correctness (binary response variable). The two-level factor *match* was added as fixed effect. For random effects, we included random intercepts for *emotion*, *participant* (nested under *decoder group*) and *speaker* (nested under *encoder group*). Adding random slopes to the model resulted in a singular fit and was therefore omitted. *Item* was not included as a crossed random effect for the same reason. The final model showed a significant main effect of *match* on *correctness* (β=0.26, SD=0.12, z=2.14, p=0.03). There was a higher proportion of correct answers if speaker and perceiver were members of the same dialect group (M_{corr}=68.7%) than if they were members of different dialect groups (M_{corr}=62.9%).

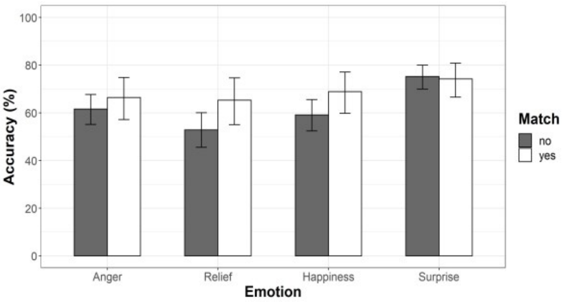

Figure 2: A*ccuracy rates of observations with matching and mismatching group affiliation for all four emotions. Error-bars represent the 95%- CI.*

Finally, we analyzed the Likert ratings that indicated the certainty of the participants answer selection by fitting a cumulative link mixed model (CLMM) using the *ordinal* R-package [24]. The Likert ratings were converted into an ordered factor and included in the model as the dependent variable. As in the GLMM, the *match* variable was added to the model as a fixed effect. The random effects structure of this model was identical to that of the GLMM except that *item* was additionally included as a random effect factor. The CLMM also showed a significant main effect of *match* on the dependent variable (β=0.32, SD=0.14, z=2.38, p=0.02). The model predicted that the probability of the highest increment on the Likert scale (5=very sure) to be chosen was higher if speaker and perceiver were members of the same variety group (*P*=0.40) than if they were members of different variety groups (*P*=0.33). On the other hand, the predicted probabilities for the four lower increments were higher for observations with mismatching group affiliation than for observations with matching group affiliation.

2.2.2. Prosodic Analysis

The 96 stimuli that were included in the statistical analysis of the perception experiment were prosodically analyzed for eleven acoustic parameters related to frequency (normalized F_0 mean and normalized F_0 range), intensity (overall mean intensity (intM), overall intensity range (intR), and mean intensity of the accented syllable of the phrase final word (accInt)), speech tempo (syllables per second and duration of the accented syllable of the phrase final word) and voice quality (harmonics-to-noise ratio (HNR), cepstral peak prominence (CPP), H1-A3, and Hammarberg-Index). The acoustic measures were extracted using the ProsodyPro Praat script [25], as well as custom Praat scripts (for global measures: F_0 mean, F_0 max, F_0 min, F_0 range, intM, intR, speech rate). For each of the eleven parameters a separate linear mixed model (LMM) was fitted using the *lmerTest* R-package [22]. The models included the acoustic measure as the dependent variable and *emotion* and *variety* as fixed effects (interaction terms). An initial model always included *speaker* as crossed random effect. Furthermore, *emotion* was included as a random slope if this did not result in a singular fit and improved the model (models were compared using the *anova()*-function). If the model did not converge, the random effects factor was removed. For the most parsimonious model, outliers were removed to formulate the final model. P-values were extracted using the Satterthwaite approximation implemented in *lmerTest* [26].

Of the eleven final models only the three models that included intensity-related measures showed a significant (or marginally significant) interaction between *emotion* and *variety*: accInt ($F(6,76)=2.97$, $p=0.01$), intM ($F(6,76)=2.78$, $p=0.02$), and intR ($F(6,76)=2.21$, $p=0.05$). All three models included *speaker* as random intercept. Figure 3 visualizes the difference in emotions across the three varieties for the mean intensity range (intR). Here, *variety* had a marginally significant main effect for sentences of the category *relief* (VN<HH,ZH; $p=0.08$, cf. Fig 3) but not for the other three emotions (all $p>0.23$); similar patterns were found for accInt and intM.

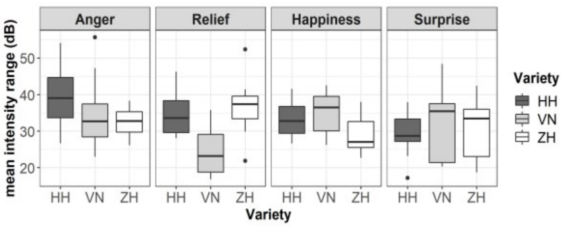

Figure 3: *Visualization of the interaction between emotion and variety for mean intensity range (dB).*

3. General Discussion

The overall main effect of *match* (i.e., matching group affiliation) on correctness of emotion judgements suggests that the in-group advantage observed for a variety of languages extends to varieties of a language: Listeners of three varieties of German recognized vocally expressed emotions better if they were encoded by speakers of the same variety. However, the magnitude and consistency of the in-group advantage numerically differed between the varieties and emotions, though not statistically. That is, of the three varieties tested here, the Zurich variety displayed the most consistent in-group advantage: Perceivers from Zurich achieved higher recognition rates than perceivers from Hamburg and Vienna for three out of the four emotions if they were encoded by speakers from Zurich. Additionally, the data collected in the Likert rating task also showed a significant main effect of matching group affiliation. This suggests that perceivers were more confident in their choice if they judged emotional expressions from their own dialect as opposed to another dialect. Considering the sense of identification and familiarity many speakers link to their own dialect, this does not come as a surprise.

Taken together, these findings highlight the role that prosody plays in the vocal communication (production and perception) of emotions. The results of the prosodic analysis of the recordings indicate that speakers of the three varieties utilize intensity related cues differently to vocally express emotions. Of the eleven acoustic measures that were inspected in the analysis, only the three measures that were related to intensity (i.e., overall mean intensity, overall intensity range, and mean intensity of the accented syllable of the phrase final word) showed a significant interaction between *variety* and *emotion*.

While the results of the study suggest an in-group advantage in emotion perception, they also show that emotions can be recognized across different varieties with relative ease (cf. [2–5, 7]). All conditions except one (VH perceivers judging *relieved* stimuli encoded by ZH speakers) displayed accuracy scores of at least two times the chance level. Our findings hence strongly support a theory of universal emotion recognition (e.g. [3, 7]). The fact that the in-group advantage prevails if the emotions are encoded by members of three culturally closely related groups suggests a secondary role of the culturally specific appraisal of emotions. Instead, speech prosody seems to be the main cause of emotional dialects.

Finally, the results of our experiment also indicate that members of certain dialect groups may be better at recognizing specific emotions. Perceivers from Zurich showed the highest accuracy scores for emotion expressions of the category surprise across all three dialects. Similarly, Viennese decoders displayed the highest accuracy scores for recordings of the category anger across all three varieties.

Note though that the sample size is rather limited, especially when subdivided into specific emotions. Therefore, claims about cross-varietal differences in the use of prosody to encode specific emotions can only reveal trends that need to be substantiated by larger studies. Additionally, prosodic differences between varieties of German have so far primarily been found on the intonational level [8–11]. Thus, conducting an analysis of the intonation of emotionally loaded stimuli may give further insights into emotional prosody in varieties of German. Future work also needs to consider (1) the status of speakers and (2) segmental differences across dialects. (1) Our study used professional speakers (actors) as encoders. It has previously been shown that acted emotion expressions differ from authentic emotion expressions both acoustically and perceptually [27]. Hence, future studies with more naturally recorded emotional expression need to corroborate our findings. Additionally, (2) our data suggest that the in-group advantage surfaces most consistently across emotions in the variety that had the largest segmental differences compared to the other two dialects (i.e., Zurich German). However, more systematic investigations are needed to draw conclusions from this pattern.

4. Conclusions

The present study provides evidence that the emotional in-group advantage observed across languages and cultures prevails even if the emotions are encoded by speakers of different varieties of the same language. The effect seems to be related to differences in the use of speech prosody to encode emotions, with intensity related cues differing most strongly across German varieties and hence contributing to the in-group advantage.

5. Acknowledgements

We thank the speakers for lending their voices. We also thank everyone who helped with recruiting listeners in the three varieties. Finally, we thank the members of the MA research colloquium (WS20/21, Carsten Eulitz) for helpful comments on the study.

6. References

[1] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis," *Psychological Bulletin*, vol. 128, no. 2, pp. 203–235, 2002, doi: 10.1037//0033-2909.128.2.203.

[2] D. C. Albas, K. W. McCluskey, and C. A. Albas, "Perception of the emotional content of speech: A comparison of two Canadian groups," *Journal of Cross-Cultural Psychology*, vol. 7, no. 4, pp. 481–490, 1976.

[3] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing Emotions in a Foreign Language," *Journal of Nonverbal Behavior*, vol. 33, no. 2, pp. 107–120, 2009, doi: 10.1007/s10919-008-0065-7.

[4] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures," *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, pp. 76–92, 2001, doi: 10.1177/0022022101032001009.

[5] W. F. Thompson and L.-L. Balkwill, "Decoding speech prosody in five languages," *Semiotica*, vol. 2006, no. 158, pp. 407–424, 2006.

[6] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, no. 4, pp. 417–435, 2009.

[7] P. Laukka *et al.,* "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features," *Journal of Personality and Social Psychology*, vol. 111, no. 5, pp. 686–705, 2016, doi: 10.1037/pspi0000066.

[8] M. Atterer and R. D. Ladd, "On the phonetics and phonology of "segmental anchoring" of F0: evidence from German," *Journal of Phonetics*, vol. 32, no. 2, pp. 177–197, 2004, doi: 10.1016/S0095-4470(03)00039-1.

[9] S. Baumann and T. Rathcke, "Disambiguating the Scope of Negation by Prosodic Cues in three Varieties of German," *Lingua*, vol. 131, pp. 29–48, 2013.

[10] B. Braun, "Effects of dialect and context in the realisation of German prenuclear accents," in *XVIth International Congress of Phonetic Sciences : ICPhS XVI, 6 - 10 August, Saarbrücken, Germany*, 2007, pp. 961–964.

[11] D. Mücke, M. Grice, A. Hermes, and J. Becker, "Prenuclear Rises in Northern and Southern German," in *4th International Conference on Speech Prosody*, 2008, pp. 245–248.

[12] C. Ulbrich, *Phonetische Untersuchungen zur Prosdie der Standardvarietaeten des Deutschen in der Bundesrepublik Deutschland, in der Schweiz und in Österreich*. Frankfurt am Main: Peter Lang, 2005.

[13] F. Schaeffler and R. Summers, "Recognizing German Dialects by Prosodic Features alone," in *Proc. ICPhS*, 1999, pp. 2311–2314.

[14] A. Leemann and B. Siebenhaar, "Perception of Dialectal Prosody," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[15] A. Leemann and B. Siebenhaar, "Intonational and temporal features of Swiss German," *Proceeding of ICPHS XVI, Saarbrücken, Germany*, pp. 957–960, 2007.

[16] J. Fleischer and S. Schmid, "Zurich German," *Journal of the International Phonetic Association*, vol. 36, no. 2, pp. 243–253, 2006.

[17] J. Fitzpatrick-Cole, "The alpine intonation of Bern Swiss German," in *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999, pp. 941–944.

[18] A. Leemann and B. Siebenhaar, "Swiss Alpine and Midland Intonation," in *Proceedings of Speech Prosody*, 2008, 289 - 282.

[19] D. J. Leiner, *SoSci Survey (Version 3.1.06).* [Online]. Available: https://www.soscisurvey.de/

[20] P. Boersma and D. Weenink, *Praat: Doing Phonetics by Computer (Version 6.1.08).* [Online]. Available: http://www.praat.org/

[21] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our Midst: An Online Behavioral Experiment Builder," *Behavior Research Methods*, vol. 52, no. 1, pp. 388–407, 2020. [Online]. Available: https://www.gorilla.sc/

[22] D. M. 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, doi: 10.18637/jss.v067.i01.

[23] R Core Team, *R: A Language and Environment for Statistical Computing (Version 4.0.3)*. Vienna, Austria. [Online]. Available: http://www.r-project.org/

[24] R. H. B. Christensen, *ordinal - Regression Models for Ordinal Data*.

[25] Y. Xu, "ProsodyPro—A tool for large-scale systematic prosody analysis," in 2013.

[26] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017, doi: 10.18637/jss.v082.i13.

[27] A. Anikin and C. F. Lima, "Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations," *Quarterly journal of experimental psychology (2006)*, vol. 71, no. 3, pp. 622–641, 2018, doi: 10.1080/17470218.2016.1270976.