Cue interaction in the perception of prosodic prominence: the role of voice quality

Bogdan Ludusan[1], *Petra Wagner*[1], *Marcin Włodarczak*[2]

[1]Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Germany
[2]Department of Linguistics, Stockholm University, Sweden
{bogdan.ludusan, petra.wagner}@uni-bielefeld.de, wlodarczak@ling.su.se

Abstract

Voice quality is an important dimension in human communication, used to mark a variety of phenomena in speech, including prosodic prominence. Even though numerous studies have shown that speakers modify their voice quality parameters for marking prosodic prominence, the impact of these modifications on perceived prominence is less studied. Our investigation looks at the effect of a well-known measure of voice quality, cepstral peak prominence (CPP), on syllabic prominence ratings given by both naive and expert listeners. Employing read speech materials in German, we quantify the role of CPP alone and in combination with other acoustic cues marking prominence, namely intensity, duration and fundamental frequency. While CPP, by itself, had a significant effect on the perceived prominence for most of the listeners, when used in conjunction with the other cues, its impact was reduced. Moreover, when assessing the importance of each of these four cues for determining the perceived prominence score we found important individual variation, as well as differences between naive and expert listeners.

Index Terms: prosodic prominence, voice quality, cepstral peak prominence, naive listeners, acoustic cues

1. Introduction

Voice quality is an important dimension in human communication, used to convey a number of characteristics, including speaker attitude, relationship to the interlocutor, or information on speech act types [1, 2], being even considered by some researchers as the fourth prosodic dimension [2], alongside intensity, duration and fundamental frequency (f0).

The role of voice quality has been examined also in relation to prosodic prominence, in particular in the marking of prominent syllables/vowels (e.g. [3, 4, 5, 6]). The findings of these studies revealed that a number of voice source characteristics (e.g. open quotient, speed quotient) vary with the prominence status of the vowel. Moreover, it has been shown that prominent vowels have a more periodic phonation than their non-prominent counterparts (e.g. [6]). Several acoustic correlates of voice quality have been investigated for prosodic prominence, including amplitude differences between various frequency bands or fundamental frequency harmonics, or the amplitude of specific frequencies in the speech spectrum or cepstrum (e.g. [7, 8, 9, 10, 11]). These measures have been found to be used for marking prominence to a certain degree, across a variety of languages.

Previous literature has established that speakers change their voice quality characteristics to mark prosodic prominence, but do listeners take advantage of these cues in the perception of prominence? It has been shown that voice quality plays a role in the perception of various linguistic quantities (e.g. pitch [12, 13]) and, thus, one would expect such an effect also for prominence. A number of studies investigating the perception of prosodic prominence (e.g. [14, 15, 16, 17]) confirmed that the main acoustic-prosodic cues: intensity, duration and f0, play a role in perception. By contrast, the use of voice quality cues in the perception of prominence by listeners has received considerably less attention. Analyzing acoustic correlates of perceived prominence in German, [18] found significant correlations between syllabic prominence and harmonics-to-noise-ratio (HNR), an acoustic measure of voice quality (although the authors explained this result by means of differences in syllabic sonority). Furthermore, even though a linear regression analysis revealed that HNR had a significant effect on prominence, this did not apply to all the considered focus conditions. The perception of prominence was examined also by [19], using synthesized stimuli that varied a global waveshape parameter, derived from a number of voice quality characteristics. The experiment found that the manipulation of those characteristics may change the degree of perceived prominence, with the findings being modulated by the position of the syllable in the sentence.

In this work, we investigate the role that voice quality plays in the perception of prominence. Differently from previous studies, we include natural speech stimuli and we do not limit our analysis to prosodic prominence due to focus, as we consider here both word and sentence stress. We chose cepstral peak prominence (CPP) as our voice quality measure, as it has been established to be the best signal-based correlate for perceived voice quality in continuous speech [20]. We then obtained prominence ratings from naive and expert listeners and we evaluated the role CPP plays in the perception of prominence. More importantly, we did not examine the effect of CPP only by itself, but also in relation to other acoustic cues shown to play a role in the perception of prominence (intensity, duration and f0).

2. Methods and materials

2.1. Dataset

We used materials from the Bonn Prosodic Database [21], a corpus consisting of German read sentences uttered by three professional speakers, in the standard Northern German variety. The sentences were orthographically transcribed and manually annotated at the segmental level and for prosodic prominence. Three expert phoneticians annotated the prominence level of each syllable using a graphical scale ranging from 0 to 31 (similarly to [22]) and high correlation coefficients (ranging from 0.74 to 0.86) were obtained between the prominence

scores given by the annotators. Each syllable was assigned the median prominence score across the three annotators, further called expert prominence score.

We then selected 70 sentences (the same 20 from the three speakers plus 10 other, sampled from the three speakers) and used them in an annotation experiment involving naive listeners (for more details, see [23]). All participants were speakers of the standard (or near-standard) Northern German variety. They were asked to listen to each sentence and to drum the sentence they heard using an electronic drum pad, using one beat for each syllable they hear. From the ten naive listeners (N1-N10; 7 females, 3 males) that participated in the drum-based annotation process, we had to exclude one (N6; female), which misunderstood the task and drummed all syllables at maximum (or near-maximum) intensity. The velocity (impact force) of the drum beats was used as operationalization for the perceptual prominence level of the naive speakers and has been shown to correlate well with three other established annotation methods (including the expert annotation provided with the corpus) [23]. We considered both the subset of 60 sentences (20 from each speaker) and the remaining subset of 10 sentences (which was used as a training phase in the experiment). We then checked the sentences annotated by each naive listener and removed the ones in which the number of drummed beats did not equal the number of syllables in the sentence. This resulted in a dataset consisting of between 537 and 587 syllables per naive listener (out of a total of 587 syllables from our 70 sentences).

Besides the individual prominence scores, given by each naive listener, we also considered an overall naive prominence score. It was computed in a similar fashion to the expert prominence score, by taking for each syllable the median value across the prominence scores given by the naive listeners for that syllable. We, thus, employed in our study three types of prominence scores: individual scores (one for each of the nine naive listeners included in the analysis), a naive median score and an expert median score. Each of these scores were z-normalized on a per-sentence basis.

2.2. Analyses

As correlate of voice quality we employed the cepstral peak prominence, as proposed by Hillenbrand and colleagues [24]. It is defined as the amplitude of the cepstral peak relative to the regression line over the entire cepstrum. CPP represents a measure of periodicity and magnitude of the harmonics above the noise level, and exhibits lower values for noisier (less periodic) signals. Thus, we expect prominent syllables to feature higher CPP values than non-prominent syllables.

For all the sentences included in our study we extracted, using the VoiceSauce toolkit [25] the following features: cepstral peak prominence (*cpp*), root-mean-square energy (*rms*, as correlate of speech intensity) and the fundamental frequency (*f0*, in Hz). They were computed using default parameters, and with a time step of 1 millisecond. From the annotations supplied with the corpus, we derived the duration (*dur*, in milliseconds) of each syllable nucleus (vowel or syllabic /n/, [26]). Then, for the features extracted with VoiceSauce, we computed their average value within each syllable nucleus in the dataset. Finally, we z-normalized each feature as follows: *rms* and *cpp* on a per-sentence basis, *f0* on a per-speaker basis and *dur* on a per-speaker and on a per-syllable nucleus category basis. The normalization of the acoustic features allowed us to compare values across different speakers and categories.

In order to test the effect of CPP on the perceived prominence scores, we fitted separate linear regression models with the prominence score (9 individual scores, 2 median scores) as dependent variable and *cpp* as predictor. For observing the role of CPP in conjunction with the other three acoustic measures we tested all combinations of the three acoustic features: by themselves (*rms*, *dur*, *f0*), in combinations of two (*rms+dur*, *rms+f0*, *dur+f0*), and all three of them (*rms+dur+f0*). We fitted, for each score, a regression model without and a model including *cpp* as predictor. We then compared the adjusted R-squared of the two models and performed an ANOVA analysis comparing the two models, in order to determine whether the addition of *cpp* brings an improvement to the baseline model (not including *cpp*).

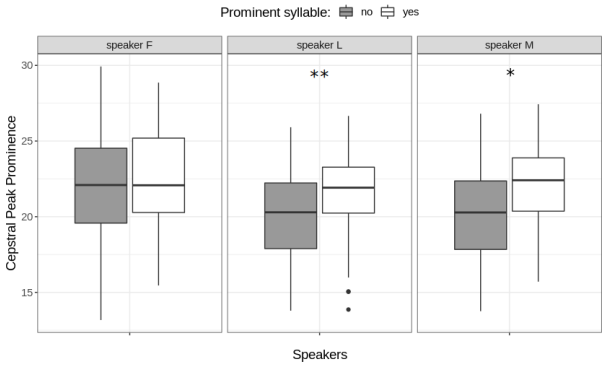

Figure 1: *Mean CPP values for the prominent and non-prominent syllables, for each speaker in our dataset. A prominent syllable was defined as being a syllable having a median expert prominence score greater or equal to the third quartile in the given sentence.*

While the regression analysis can provide us with information on whether a specific features has a significant effect on the prominence score, it cannot give us an intuitive ranking of the importance of each feature in determining the perceived prominence score. For this, we employed a Random Forest [27] regression analysis, by which a model consisting of 500 trees was used to learn each prominence score, considering the four cues as input features. The importance of each cue was determined by computing the decrease in node impurity from splitting on that cue, averaged over all trees in the model. The decrease in node impurity was measured by means of the residual sum of squares. The Random Forest regression was run for 100 times, for each score, and the sum of the importance across all the runs computed. Then, for each cue, its importance was divided by the total importance, across all four cues, thus obtaining values between 0 and 1, which can be compared between the different prominence scores.

All statistical analyses were performed using the appropriate functions provided by the R software [28], including the randomForest package [29] for performing Random Forest regression and feature importance analysis.

3. Results

Previous work has shown that speakers may employ different strategies for the marking of prosodic prominence by means of voice quality changes [30]. Therefore, we first performed an analysis of the data produced by the speakers in our corpus to confirm the effect of CPP on the production of prominent syllables in the analyzed materials. For this, we made

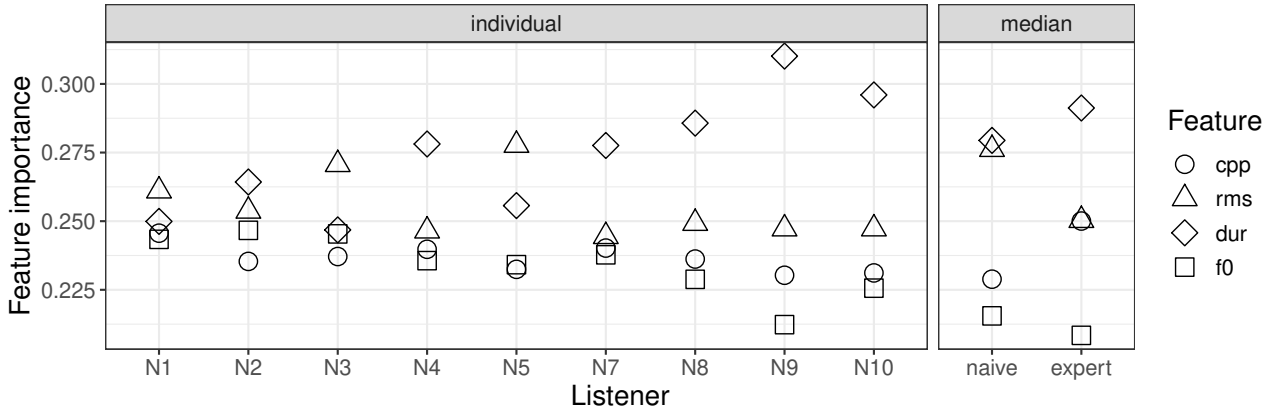

Figure 2: *Individual feature importance, as given by the Random Forest analysis, for each naive listener (N1-N10), as well as for the median prominence scores across naive and expert listeners, respectively. The sum of the importance of all four cues, per speaker, equals 1.*

use of the prominence annotations supplied with the dataset (expert score). For every sentence, we considered a syllable to be prominent if its expert prominence score was greater or equal to the third quartile of all syllable prominence scores of that sentence, otherwise it was considered non-prominent. Figure 1 illustrates the mean CPP values for the prominent and the non-prominent syllables, for the three speakers in our corpus, with higher values obtained for the former category. A Wilcoxon signed rank test was used to test the difference in CPP values between the two cases (each sentence gave two values: one for prominent and one for non-prominent syllables therein). Although a significant difference was observed when the entire dataset was included ($p = 5.5e^{-5}$), when looking at each speaker, individually, the difference reached significance only for two of the speakers (speaker F: $p = 0.317$, speaker L: $p = 1.1e^{-3}$ and speaker M: $p = 0.012$). As one of the speakers did not produce consistent voice quality changes captured by our measure, CPP, listeners will have no such information to use for the perception of prominence. We, thus, present in this section the results obtained based on the data produced by speakers L and M, with any differences between this set and the whole corpus being discussed in the next section.

The results of the linear regression analyses are reported in Table 1. We see that *cpp* has a significant effect on the prominence rating of 7 (out of 9) naive listeners, as well as on the median prominence scores, both naive and expert (first line of the table). When examining its effect in combination with the other acoustic cues, we observed a significant improvement for the models that consider *cpp* as predictor (compared to the ones without *cpp*) for all single-cue models, as well as for all multiple-cue models not including both *rms* and *dur* (lines 2-8). Out of the 7 naive listeners that seem to use *cpp* for their prominence perception, the majority of them continue to use the cue also when other cues are taken into account, except for when both *rms* and *dur* are involved (*rms+dur* and *rms+dur+f0*, lines 2-8, column Individual). Moreover, better discrimination was obtained when including *cpp* as independent variable for all the models predicting the expert prominence score and for all naive median score models, again except for those which consider both *rms* and *dur* (lines 2-8, columns Naive and Expert). In order to better understand the reduced effect of *cpp* in the models employing both *rms* and *dur*, we looked at the

Table 1: *Linear regression analysis results for each individual naive listener score, as well as for the median perceived prominence across naive and expert listeners, respectively. Column Individual illustrates the number of naive listeners for which a significant effect of CPP was observed. The last two columns show the p-values obtained from the analyses. The fist line represents the model containing only CPP as predictor. The following lines display the ANOVA comparison between a model employing the listed independent variables and a model using those predictors and CPP.*

Predictors	**Individual**	**Median**	
		Naive	**Expert**
cpp	7/9	$5.3e^{-5}$	$2.6e^{-9}$
rms	4/9	.021	$6.5e^{-6}$
dur	6/9	$6.8e^{-4}$	$1.2e^{-7}$
f0	7/9	$6.4e^{-5}$	$3.0e^{-9}$
rms+dur	1/9	.256	$9.3e^{-4}$
rms+f0	5/9	.018	$6.4e^{-6}$
dur+f0	6/9	$1.3e^{-3}$	$2.3e^{-7}$
rms+dur+f0	1/9	.260	$9.4e^{-4}$

Spearman correlation between the former and the latter two predictors. It revealed significant correlations between *cpp* and *rms*: $\rho = .368, p = 1.1e^{-13}$, and between *cpp* and *dur*: $\rho = .191, p = 1.7e^{-4}$ (compared to a lack of correlation between *cpp* and *f0*: $\rho = .051, p = .32$), which may, at least partly, explain this reduced effect.

Looking at the importance of the investigated cues, as given by a Random Forest regression model, we see important variation across the naive listeners, both in the ranking, as well as in the importance of the individual cues. Some listeners (N1-N2) tend to give rather similar weights to the four cues, while others (N8-N10) show a strong preference for one of the cues (*dur*, in this case). With regard to the obtained ranking for individual prominence scores, *dur* and *rms* seem to be the most important cues (for 6 and 3 listeners, respectively), while *f0* and *cpp* were found to be the least important ones (in 6 and 3 cases, respectively). As for the median prominence scores, the results show that although the weighting of the cues differs between

the two cases (naive/expert), their ranking is the same, with the listeners basing their judgments on *dur*, followed by *rms*, *cpp* and, lastly, *f0*. Furthermore, while naive listeners give quite different weights to *cpp* and *rms*, expert listeners give them a similar importance. Taking a closer look at the cue importance for median scores, one can see that the biggest differences are for these two cues, with the importance of *cpp* increasing by 2.1% and that of *rms* decreasing by 2.6% for the expert score compared to the naive one. The cue importance differences for the individual and median prominence scores were tested using Wilcoxon signed rank tests. All differences between cues importance, except for *cpp-rms* in the case of the expert score, were found to be significant. We then checked whether the individual variation we observed in the naive listeners data may be explained by gender differences. For this, we tested whether the gender of the listener had an effect on the importance of each of the four acoustic measures, by performing individual Kruskal-Wallis tests. As none of the tests revealed a significant effect, we can conclude that the individual variation cannot be explained by differences in gender.

4. Discussion and conclusions

The results of our study are in line with those of previous work. Similarly to [18, 19], we observed an effect of voice quality on the rating of perceptual prominence. Our investigation extended these findings by showing that both naive and expert listeners make use of voice quality information to perceive prominence, although they give it different weights. The individual variation in using voice quality in the production of prominence reported by [30] was found also here, with one of the three speakers in our dataset not marking prominence consistently by means of CPP. Moreover, our results indicate that individual variation is present not only in the production, but also in the perception of prominence, based on the fact that not all of our naive listeners make use of voice quality information.

Different from prior studies on the effect of voice quality cues on prominence, we did not employ highly controlled materials, in which individual words or balanced prosodic focus structures were compared. Rather, we had a mixed set of individual sentences, varying in length, content and pragmatic expression. Despite this (comparative) lack of control, the effect of voice quality on prominence perception remained stable, but subtle. Nonetheless, we concede that our material still consisted of read isolated sentences produced under laboratory conditions. For the purpose of studying the impact of voice quality on prominence, we believe this might have been, actually, detrimental: in spontaneous interaction, we expect speakers and listeners to be forced to exploit voice quality cues more strongly, as various factors may interfere with the production of other cues that mark prosodic prominence. For instance, loud speech (e.g., due to noisy conditions) may neutralize intensity, while creak (e.g., due to utterance planning or turn taking) may neutralize pitch. Therefore, an extension of these findings to spontaneous speech would be an appropriate next step, for a better understanding of the role of voice quality in prosodic prominence expression.

As mentioned in the previous section, we reported the analyses based on data produced by the two speakers that showed significant differences with respect to CPP, between prominent and non-prominent syllables. Performing the same analyses, but using the full dataset of three speakers, we obtained similar findings for the fitted linear models. Differences in significance were found only for three of the models fitted with the individual scores and one median naive score model. The results for the cue importance, as given by the Random Forest regression, showed a lower importance for CPP than for the fundamental frequency. Also for the median scores, CPP exhibited a lower importance than it was determined on the dataset based on the two speaker, with CPP being the least important cue for the naive score and the second least important one for the expert score. These differences were to be expected, as the speaker that was removed did not mark prominence with voice quality changes captured by CPP.

To summarize, we have seen that voice quality information, in the form of CPP, may be used by listeners to discriminate prominent from non-prominent syllables. However, not every naive listener employed the information given by CPP in the perception of prominence and this information did not bring much improvement over models taking into account both syllable nucleus intensity and duration. Nevertheless, an interesting difference was observed between naive and experts listeners, with the later group giving a higher importance to voice quality information than the former group. Moreover, all the models fitted with the expert prominence score showed improvements when CPP was added to the model. It is encouraging to see that most of the models including CPP did bring an improvement over those without CPP and that the its overall importance is similar or higher than that of other, better studied, cues. As we expect to see a higher importance of voice quality in spontaneous materials we intend to expand our investigation with a perception experiment involving conversational data.

5. Acknowledgements

The work was funded by Swedish Research Council project 2019-02932 *Prosodiska funktioner hos röstkvalitetsdynamik* (*Prosodic functions of voice quality dynamics*) to Marcin Włodarczak.

6. References

[1] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[2] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *Proc. of ICPhS*, 2003, pp. 2417–2420.

[3] J. Koreman, "The effects of stress and F0 on the voice source," *Phonus*, vol. 1, pp. 105–120, 1995.

[4] C. Mooshammer, "Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 1047–1058, 2010.

[5] M. Vainio, M. Airas, J. Järvikivi, and P. Alku, "Laryngeal voice quality in the expression of focus," in *Proc. of INTERSPEECH*, 2010, pp. 921–924.

[6] L. Lancia, D. Voigt, and G. Krasovitskiy, "Characterization of laryngealization as irregular vocal fold vibration and interaction with prosodic prominence," *Journal of Phonetics*, vol. 54, pp. 80–97, 2016.

[7] A. Sluijter and V. van Heuven, "Acoustic correlates of linguistic stress and accent in Dutch and American English," in *Proc. of ICSLP*, vol. 2, 1996, pp. 630–633.

[8] N. Campbell and M. Beckman, "Stress, prominence, and spectral tilt," in *Intonation: Theory, Models and Applications*, 1997, pp. 67–70.

[9] P. Prieto and M. Ortega-Llebaria, "Stress and accent in Catalan and Spanish: Patterns of duration, vowel quality, overall intensity, and spectral balance," in *Proc. of Speech Prosody*, 2006, pp. 337–340.

[10] M. Garellek and J. White, "Phonetics of Tongan stress," *Journal of the International Phonetic Association*, vol. 45, no. 1, pp. 13–34, 2015.

[11] S. Kakouros, O. Räsänen, and P. Alku, "Evaluation of spectral tilt measures for sentence prominence under different noise conditions," in *Proc. of INTERSPEECH*, 2017, pp. 3211–3215.

[12] M. P. Bissiri and M. Zellers, "Perception of pitch in glottalizations of varying duration by German listeners," in *Proc. of ICPhS*, 2015, p. 691.

[13] J. Kuang and M. Liberman, "Integrating voice quality cues in the pitch perception of speech and non-speech utterances," *Frontiers in Psychology*, vol. 9, p. 2147, 2018.

[14] J. Pierrehumbert, "The perception of fundamental frequency declination," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 363–369, 1979.

[15] C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump, and J. Terken, "The perceptual prominence of fundamental frequency peaks," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 3009–3022, 1997.

[16] M. Vainio and J. Järvikivi, "Tonal features, intensity, and word order in the perception of prominence," *Journal of Phonetics*, vol. 34, no. 3, pp. 319–342, 2006.

[17] J. Bishop, G. Kuo, and B. Kim, "Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription," *Journal of Phonetics*, vol. 82, p. 100977, 2020.

[18] H. Mixdorff, C. Cossio-Mercado, A. Hönemann, J. Gurlekian, D. Evin, and H. Torres, "Acoustic correlates of perceived syllable prominence in German," in *Proc. of INTERSPEECH*, 2015, pp. 51–55.

[19] I. Yanushevskaya, A. Murphy, C. Gobl, and A. Ní Chasaide, "Perceptual salience of voice source parameters in signaling focal prominence," in *Proc. of INTERSPEECH*, 2016, pp. 3161–3165.

[20] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2619–2634, 2009.

[21] B. Heuft, T. Portele, C. Widera, P. Wagner, and M. Wolters, "Perceptual prominence," in *Speech and Signals – Aspects of Speech Synthesis and Automatic Speech Recognition*, W. Sendlmeier, Ed. Frankfurt am Main: Hektor, 2000, pp. 97–116.

[22] G. Fant and A. Kruckenberg, "Preliminaries to the study of swedish prose reading and reading style," *STL-QPSR*, vol. 2, no. 1989, pp. 1–83, 1989.

[23] 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, p. 100911, 2019.

[24] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[25] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proc. of ICPhS*, 2011, pp. 1846–1849.

[26] K. Kohler, "German," *Journal of the International Phonetic Association*, vol. 20, no. 1, p. 48–50, 1990.

[27] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[29] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[30] I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Cross-speaker variation in voice source correlates of focus and deaccentuation," in *Proc. of INTERSPEECH*, 2017, pp. 1034–1038.