INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The Role of Voice Quality in the Perception of Prominence in Synthetic Speech

Andy Murphy, Irena Yanushevskaya, Ailbhe Ní Chasaide, Christer Gobl

Phonetics and Speech Laboratory, Trinity College Dublin
murpha61@tcd.ie, yanushei@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

Abstract

This paper explores how prominence can be modelled in speech synthesis through voice quality variation. Synthetic utterances varying in voice quality (breathy, modal, tense) were generated using a glottal source model where the global waveshape parameter R_d was the main control parameter and f_0 was not varied. A manipulation task perception experiment was conducted to establish perceptually salient R_d values in the signalling of focus. The participants were presented with mini-dialogues designed to elicit narrow focus (with different focal syllable locations) and were asked to manipulate an unknown parameter in the synthetic utterances to produce a natural response. The results showed that participants manipulated R_d not only in focal syllables, but also in the pre- and postfocal material. The direction of R_d manipulation in the focal syllables was the same across the three voice qualities – towards decreased R_d values (tenser phonation). The magnitude of the decrease in R_d was significantly less for tense voice compared to breathy and modal voice, but did not vary with the location of the focal syllable in the utterance. Overall, the results suggest that R_d is effective as a control parameter for modelling prominence in synthetic speech.

Index Terms: global waveshape parameter R_d, speech synthesis, voice quality, perception test, prominence, manipulation task

1. Introduction

This paper is part of ongoing work to explore how prominence in synthetic speech can be modelled using a limited set of voice source parameters. This research can be directly integrated into the synthetic voices currently being developed for Irish dialects (www.abair.ie [1]) which are being used in interactive games and applications for language learning [2, 3]. It is important also because it furthers understanding of linguistic prosody, e.g., the interaction of properties of the source in the signalling of linguistic prominence.

When the goal is to reduce the number of control parameters in synthesis, the global waveshape parameter R_d [4, 5] has shown promising results. [6] describes how it can be used to perform breathiness and pitch transformations. Both [7] and [8] have implemented unit selection speech synthesis systems with voice modification capabilities using R_d as a control parameter.

The findings in [9, 10] suggest that manipulating R_d without f_0 salience can be used in synthesis to generate focal prominence and emotional colouring. [11] explored optimal implementations of R_d. Listening tests were conducted in these earlier papers in which synthesised utterances were presented to the participants who were asked to assess the degree of prominence of the syllables in the utterances, or the kind and degree of affective colouring. The stimuli were synthesised and parameters varied based on earlier analytical studies and the ranges and extent of variation were defined by the experimenters.

Our studies on R_d manipulation in signalling prominence used the R_d settings for modal voice but have not considered other phonatory settings. Earlier analytical studies exploring cross-speaker variation in the signalling of focus [12] and source correlates of focal prominence across different voice qualities [13, 14] showed that baseline (speaker-specific, habitual) voice quality is likely to have an impact on speaker strategies in signalling focus. Given these findings, it is important to establish the effect of differences in voice quality on signalling prominence in synthetic speech. Earlier studies have also shown that different degrees of prominence are associated with syllables located in different parts of the utterance [9]. This effect was attributed to the lack of f_0 manipulation applied to the stimuli.

The aim of this study is to experimentally obtain perceptually salient settings of R_d for signalling focal prominence for three voice qualities (breathy, modal and tense). An approach different from previous studies is proposed: rather than rating perceptual salience of ready-made stimuli, participants actively manipulate the R_d parameter to generate perceived prominence. The methodology is similar (though not identical) to the adjustment task, e.g., [15] where the listeners were asked to adjust jitter, shimmer and harmonics-to-noise ratio in synthetic stimuli to match natural voice samples or [16] which described a quasi-adjustment task involving judgment of the relative prominence of f_0 peaks. A manipulation task rather than an adjustment task was used, that is, the participants were not given a naturally produced example sentence to match. Instead, they were presented with mini-dialogues constructed to elicit narrow focus and were asked to manipulate utterances synthesised with different voice qualities so that they sounded acceptable/natural as a response to a given question. The objective is to establish what R_d values are salient for different phonation types and what degree of R_d salience is required to make a particular syllable prominent.

By giving subjects direct control over acoustic variables, one is obtaining a (hopefully) more accurate representation of what they perceive. This approach acts as a testing platform for ongoing research into a user-controlled text-to-speech (TTS) synthesis interface being developed as part of the ABAIR project [1].

2. Material and method

2.1. Baseline sentences

The stimuli for the perception test were based on a recording of an all-voiced sentence ‘We were away a day ago’ spoken by a male Irish English speaker. The vowel quality in the potentially accentable syllables WAY and DAY was the same; in

http://dx.doi.org/10.21437/Interspeech.2019-2761

the original recording the duration of the vowels in these syllables was approximately the same (162 ms and 170 ms respectively).

The utterance was automatically inverse filtered using the iterative adaptive inverse filtering (IAIF) approach [17] to obtain an estimate of the differentiated glottal source. Voice source parameterisation was then carried out using the Liljencrants-Fant (LF) model [18] according to the dynamic programming method described in [19]. A synthetic source signal was created by concatenating LF pulses generated using a global waveshape parameter R_d [4, 5] (described in more detail in section 2.2) and adding amplitude modulated aspiration noise [20]. The source signal was then passed through a vocal tract filter derived from the coefficients obtained from IAIF.

Three baseline sentences were created representing breathy, modal, and tense voice quality. The R_d values in each of these sentences were kept constant and were set to 1.6 for breathy, 1 for modal, and 0.7 for tense voice (based on the production data in [13] and auditory analysis). f_0 was set to its average value (104 Hz) with the addition of a degree of declination (8.5 Hz/s), and was kept the same in all three sentences. Duration was not manipulated.

Preliminary auditory analysis by the authors confirmed that there was no prominence on the potentially accentable syllables in the resulting sentences; in other words, they were both equally non-prominent.

2.2. Implementation of R_d variation

R_d is used as the only control parameter in this experiment. The R_d parameter is derived from f_0, E_e and U_p as follows: $(1/0.11)\times(f_0 \cdot U_p/E_e)$, where E_e is the excitation strength (measured as the negative amplitude of the differentiated glottal flow at the time point of maximum waveform discontinuity) and U_p is the peak flow of the glottal pulse (see Figure 1). Variation in R_d is proposed to reflect voice source variation along the tense-lax continuum; the values typically range between 0.3 (tense voice) to 2.5 (breathy voice).

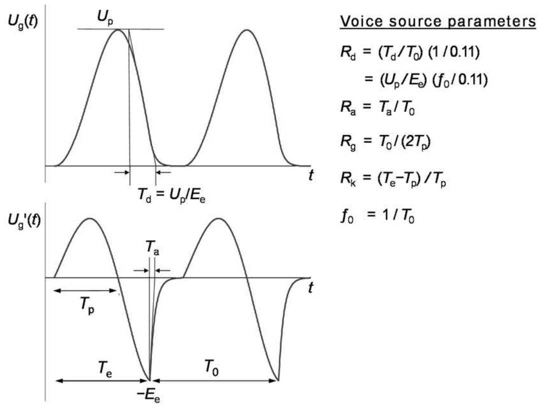


Figure 1: *Parameters used to generate the LF model waveform (adapted from [4]). Upper panel: glottal flow; lower panel: glottal flow derivative.*

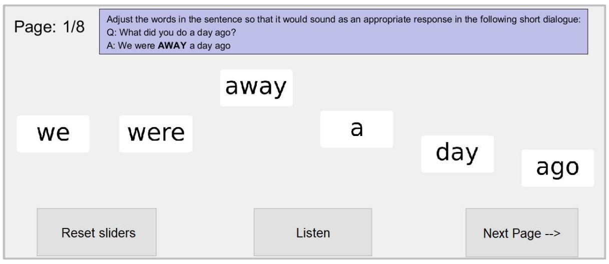
By changing R_d, other parameters of the glottal source such as R_a and R_k also vary, and these changes can be predicted from R_d. To synthesize the LF model waveform, data for the full set of LF parameters are required and were calculated from R_d using the parameter correlations presented in [4] (see also [21]).

R_d is determined by E_e, U_p and f_0 and to effect variation in R_d changes to these parameters are required. As the intention was not to vary f_0, and given the results in [11], which suggested that a more perceptually salient R_d implementation involves fixed U_p and varying E_e, these were the settings used in the current experiment. Note that U_p and R_d are not actual parameters of the LF model and cannot be directly controlled unless an iterative algorithm such as the one presented in [22] is used. However, for this experiment we used the approximations suggested in [4], which were deemed sufficiently accurate.

Word boundaries were annotated in Praat [23] along with vowel midpoints and vowel boundaries. The R_d values were kept constant at word boundaries, but were allowed to vary across the vowel segments. Vowel midpoints were used to extract R_d values obtained as a result of the listening test. R_d was allowed to vary within ranges set for each phonation type and using scaling factors that were applied across vowel segments in manipulated words. Note that in *away* and *ago* only the stressed vowels were manipulated.

2.3. Synthesis user interface

A user interface (see Figure 2) was designed for the listening test. Each word was represented by a block which, when dragged up and down, controlled a scaling factor which was then applied to the original (baseline) R_d contour. The scaling factors ranged between 0.5 and 2, so that the baseline R_d value could be halved or doubled at either end of the scale.


Page: 1/8

Adjust the words in the sentence so that it would sound as an appropriate response in the following short dialogue:
Q: What did you do a day ago?
A: We were **AWAY** a day ago

Figure 2: *User interface for the listening test.*

Constraints were also applied so that the R_d values remained inside the ranges set for each voice quality: 1-2.3 for breathy voice, 0.6-1.4 for modal voice, and 0.35-1.4 for tense voice. These values were derived from data examined in [13].

2.4. Listening test

42 participants (all native speakers of English) took part in the test. The participants were asked to manipulate the blocks/words in the utterances so that the resulting sentence could be an acceptable/natural sounding response in a mini-dialogue involving narrow focus, e.g.

Dialogue 1: Q: What did you do a day ago?
A: We were a**WAY** a day ago

Dialogue 2: Q: When were you away?
A: We were away a **DAY** ago.

There were eight utterances to manipulate. These included three instances of each of the dialogues, one for each of the three different voice qualities, in random order. Two more utterances were included at the beginning of the test to allow the participants to familiarise themselves with the procedure; the results from these were discarded. The participants were allowed to listen to the results of their manipulation and make changes as many times as they wished. The stimuli were presented through high quality closed back headphones in a quiet environment. The test took approximately 10 minutes to complete.

Hypotheses:

- Perceptually salient prominence can be generated by manipulating R_d (in the absence of f_0 variation).
- The magnitude of R_d excursions in the accented syllables are different for different baseline qualities.
- Manipulations of R_d will differ depending on the location of the focally accented syllable in the utterance.

3. Results and discussion

The R_d values at vowel midpoints were extracted from each response (42 participants x 6 utterances = 252). The average R_d contours and 95% confidence intervals of response sentences in which WAY and DAY were made prominent are shown in Figure 3. Note that the increased values of R_d correspond to laxer/breathier phonation and lower values of R_d to tenser phonation. Thus, the peaks in Figure 3 show an increase in phonatory tension.

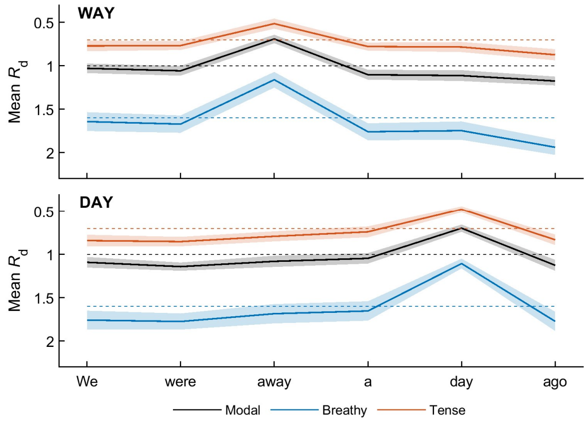


Figure 3: *Mean R_d contours for WAY and DAY responses. Baseline values are shown by dotted lines. Shaded regions indicate 95% confidence intervals.*

It is clear in Figure 3 that:

- The direction of R_d manipulation is the same across the three phonation types – towards lowering R_d /increasing tension relative to the baseline.
- The magnitude of R_d manipulation/excursion from the baseline varies with voice quality when plotted on a linear scale: the largest for breathy voice and the smallest for tense.
- R_d manipulation in pre- and post-focal material corresponds to changes towards breathier/laxer phonation relative to the original baseline.
- There is a considerable drop in tension/rise in R_d in postfocal *ago* when the focal syllable is WAY.

The magnitude of R_d manipulation/excursions in the focal syllables was measured as local 'protrusions' relative to the adjacent unaccented syllables (calculated as the difference between the focal R_d value and the average of the R_d values in the adjacent unaccented syllables and expressed as percentages relative to the ranges of R_d values) .

Linear mixed-effect model analysis was used to test if the magnitude of R_d excursion is affected significantly by the baseline voice quality (VQ) and the location of the focal syllable (Focus) in the utterance. Analyses were conducted in the R environment [24] using the *lme4* [version 1.1-20] package [25] for model fitting. The *lmerTest* package [26] was used for step-down model simplification by eliminating non-significant effects and for calculating denominator degrees of freedom using Satterthwaite's approximations. The models were fitted using the maximum likelihood (ML) method. The initial model included VQ and Focus as the main predictor variables (fixed effects) as well as their interaction; random effects included by-subject random intercepts and slopes: [Rd~VQ*Focus+(1+VQ|Participant)]. The final reduced model included VQ as the only fixed predictor and by-subject random intercept [Rd~VQ+(1|Participant)]. The location of the focal syllable (Focus) and the VQ*Focus interaction were not significant and were excluded from the model. ICC (indicative of the correlation of the items within a cluster) as well as marginal and conditional R-squared statistics [27] were obtained using the *sjPlot* package [28]. Marginal R-squared describes the proportion of the variance explained by the fixed effects; conditional R-squared indicates the variance explained by both fixed and random effects.

The summary of the estimated coefficients of the mixed effect model fitted to the R_d values (calculated as local 'protrusions' relative to the voice quality specific R_d ranges) obtained in the listening test is given in Table 1 (see also Figure 4). Based on marginal and conditional R^2 values, the amount of variance explained by the random effects amounted to about 40% of the variance. Fixed effect of baseline voice quality accounts for about 14% of the variance. Analysis of the fixed effects suggests a statistically significant association between perceptually salient R_d values in the focal syllables and baseline voice quality. While there is a drop in R_d across all voice qualities in the focal syllable, the magnitude of this drop is significantly less in tense voice compared to modal and breathy voice ($\beta = 21.62$, $p<0.001$ and $\beta = 17.73$, $p<0.001$ respectively). The magnitude of perceptually salient R_d lowering is not significantly different in modal and breathy voice. As mentioned earlier, the location of the focal syllable in the utterance (earlier or later) has no significant effect on the magnitude of R_d excursions associated with that syllable within the same voice quality type.

Table 1: *Estimated coefficients, confidence intervals and t-values for the mixed effect model fitted to obtained R_d local protrusion values.*

Mean (Intercept)	β_0	*CI*	*t*	*p*
Breathy	-44.71	-50.68 – -38.74	-14.67	<0.001
Modal	-48.60	-54.57 – -42.63	-15.95	<0.001
Tense	-26.98	-32.96 – -21.01	-8.85	<0.001
Contrasts	β_1	*CI*	*t*	*p*
breathy v.modal	-3.89	-9.07 – 1.29	-1.47	0.142
breathy v.tense	17.73	12.55 – 22.90	6.71	**<0.001**
modal v. tense	21.62	16.44 – 26.80	8.18	**<0.001**
Random effects				
ICC Participant	0.45			
Observations	252			
Marginal R^2 / Conditional R^2	0.142/ 0.531			

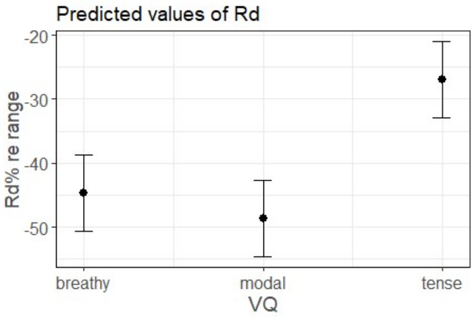


Figure 4: *Predicted values of R_d and 95% CI (local protrusions relative to adjacent unaccented syllables expressed as % relative to voice quality R_d range).*

Based on the analysis and given the initial difference in the R_d ranges across voice qualities, it appears that a higher degree of manipulation is required for modal and breathy voice than for tense voice. However, applying a logarithmic scale to the data would result in a closer representation of perceptually meaningful distances between voice qualities (see Figure 5). Mixed model analysis of the log-transformed data confirmed that the difference in the magnitude of manipulation across different voice qualities is not statistically significant. This suggests that, when using R_d as a control parameter in synthesis, logarithmic manipulations will be more appropriate.

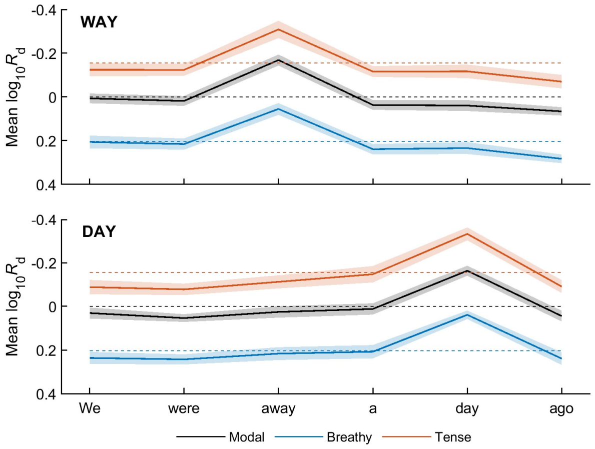


Figure 5: *Mean $\log_{10} R_d$ contours for WAY and DAY responses. Baseline values are shown by dotted lines. Shaded regions indicate 95% confidence intervals.*

The results support the initial hypothesis that perceptually salient prominence can be generated by manipulating R_d (in the absence of f_0 variation). It appears that this manipulation is not confined to the focal syllable alone but affects the pre- and post-focal material, and as a result, the overall contour of the utterance. The direction of the excursions was the same across synthetic qualities: the changes of the control parameter made by the participants in the focal syllable were all towards lowering R_d (perceptually tenser voice). Earlier studies with ready-made, predefined stimuli showed similar findings – that manipulating R_d towards tenser settings result in signalling prominence (at least for speakers of English and Irish who were participants in those earlier studies). Production data in [29] also support these findings. However, this trend may not necessarily be independent of language, as is illustrated by the results in [14] for Finnish where focal syllables were characterised by higher NAQ values (laxer/breathier phonation) [30].

The hypothesis that the magnitude of R_d excursions in the accented syllables would differ for different baseline qualities was supported when the values were normalized to the voice quality specific range and a linear scale was used. This is in keeping with the fact that the range of R_d values for laxer, breathier voice relative to modal (1-2.5) is larger than that for tense voice (0.3-1), meaning there is more room for adjustment when the R_d is in the lax/breathy range. This non-linear relationship was accounted for in [10] by log-transforming the R_d values. In this study, the differences in the magnitude of R_d excursions were not statistically significant when the values were scaled logarithmically.

Manipulations of R_d did not differ depending on the location of the focally accented syllable in the utterance. The modelling of prominence using R_d in [9, 11] was carried out by adding R_d excursions on top of an overall R_d declination, and the results of perception tests showed a significant difference in the perceptual prominence of a syllable depending on its phrase location (earlier focal syllable was perceived as more prominent than the one located later in the utterance). This trend did not emerge in the present results. The nature of task in this study was quite different from the previous studies insofar as the participants were tasked with generating an utterance with a prominent syllable rather than rating perceived prominence of predefined stimuli. It might be the case that additional R_d declination (used in [9, 11]) plays a role in this difference, and it requires further exploration.

4. Conclusions

This study involved a user-driven manipulation task experiment in an effort to obtain perceptually salient R_d parameter contours in signalling focal prominence. This type of experimental setup has not been widely used. Since the results represent the output of active manipulation of acoustic parameters, they can potentially give a more accurate picture of the extent of manipulation required. The participants chose to manipulate not only the focal syllables but the pre- and post-focal material as well. The results clearly show that it is not enough to simply scale the overall contour linearly; as would be expected, the magnitude of the focal R_d peaks needs to be scaled in proportion to the baseline R_d value. Although it has not been directly assessed in this experiment, it would be interesting to test and compare the relative importance of the extent of R_d decrease in the focal syllable (tenser phonation) and of the R_d increase (laxer phonation) in the pre-/post-focal material (deaccentuation) in signalling focal prominence.

It is hoped that this experiment will assist with the implementation of prosody-related manipulations in our Irish TTS systems, where manipulation of prominence in interactive contexts would be important. This perception tool can be used to extend our understanding of listeners' use of voice quality cues in linguistic and expressive prosody, with a view to implementation in synthesis.

5. Acknowledgements

This research is supported by the Dept. of Culture, Heritage, and the Gaeltacht (ABAIR project) and CabairE (An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta).

6. References

[1] A. Ní Chasaide, N. Ní Chiaráin, C. Wendler, H. Berthelsen, A. Murphy, and C. Gobl, "The ABAIR initiative: bringing spoken Irish into the digital space," in *Interspeech 2017*, Stockholm, Sweden, 2017, pp. 2113-2117.

[2] N. Ní Chiaráin and A. Ní Chasaide, "The Digichaint interactive game as a virtual learning environment for Irish," in *CALL communities and culture - short papers from EUROCALL 2016*, Limassol, Cyprus, 2016, pp. 330-336.

[3] N. Ní Chiaráin and A. Ní Chasaide, "Chatbot technology with synthetic voices in the acquisition of an endangered language: Motivation, development and evaluation of a platform for Irish," in *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoro, Slovenia, 2016, pp. 3429 - 3435.

[4] G. Fant, "The LF-model revisited: transformations and frequency domain analysis," *STL-QPSR,* vol. 2-3, pp. 119-156, 1995.

[5] G. Fant, "The voice source in connected speech," *Speech Communication,* vol. 22, pp. 125-139, 1997.

[6] G. Degottex, A. Roebel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5128-5131.

[7] A. Sorin, S. Shechtman, and A. Rendel, "Semi parametric concatenative TTS with instant voice modification capabilities," in *Interspeech 2017*, Stockholm, Sweden, 2017, pp. 1373-1377.

[8] C. Buchanan, M. P. Aylett, and D. Braude, "Adding personality to neutral speech synthesis voices," in *20th International Conference, SPECOM 2018, Proceedings*, A. Karpov, O. Jokisch, and R. Potapova, Eds., ed Leipzig, Germany, 2018, pp. 49-57.

[9] I. Yanushevskaya, A. Murphy, C. Gobl, and A. Ní Chasaide, "Perceptual salience of voice source parameters in signaling focal prominence," in *Interspeech 2016*, San Francisco, CA, 2016, pp. 3161-3165.

[10] A. Murphy, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Rd as a control parameter to explore affective correlates of the tense-lax continuum," in *Interspeech 2017*, Stockholm, Sweden, 2017, pp. 3916-3920.

[11] A. Murphy, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Voice source contribution to prominence perception: *Rd* implementation," in *Interspeech 2018*, Hyderabad, India, 2018, pp. 217-221.

[12] I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Cross-speaker variation in voice source correlates of focus and deaccentuation," in *Interspeech 2017*, Stockholm, Sweden, 2017, pp. 1034-1038.

[13] I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "The interaction of long-term voice quality with the realisation of focus," in *Speech Prosody 2016*, Boston, MA, 2016, pp. 1-5.

[14] M. Vainio, M. Airas, J. Järvikivi, and P. Alku, "Laryngeal voice quality in the expression of focus," in *Interspeech 2010*, Chiba, Japan, 2010, pp. 921-924.

[15] J. Kreiman and B. R. Gerratt, "Perception of aperiodicity in pathological voice," *Journal of the Acoustical Society of America,* vol. 117, pp. 2201-2211, 2005.

[16] C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump, and J. Terken, "The perceptual prominence of fundamental frequency peaks," *Journal of the Acoustical Society of America,* vol. 102, pp. 3009-3022, 1997.

[17] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication,* vol. 11, pp. 109-118, 1992.

[18] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR,* vol. 4, pp. 1-13, 1985.

[19] J. Kane and C. Gobl, "Automating manual user strategies for precise voice source analysis," *Speech Communication,* vol. 55, pp. 397-414, 2013.

[20] C. Gobl, "Modelling aspiration noise during phonation using the LF voice source model," in *Interspeech 2006*, Pittsburg, PA, USA, 2006, pp. 965-968.

[21] C. Gobl, "The Voice Source in Speech Communication: Production and Perception Experiments Involving Inverse Filtering and Synthesis," PhD thesis, KTH, Stockholm, Sweden, 2003.

[22] C. Gobl, "Reshaping the transformed LF model: generating the glottal source from the waveshape parameter R_d," in *Interspeech 2017*, Stockholm, Sweden, 2017, pp. 3008-3012.

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2017.

[24] R Core Team, "R: A language and environment for statistical computing," ed. Vienna, Austria: R Foundation for Statistical Computing, 2019.

[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software,* vol. 67, pp. 1-48, 2015.

[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, pp. 1-26, 2017.

[27] S. Nakagawa, P. C. D. Johnson, and H. Schielzeth, "The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded," *Journal of The Royal Society Interface,* vol. 14, p. 20170213, 2017/09/30 2017.

[28] D. Lüdecke, "sjPlot: Data Visualization for Statistics in Social Science," ed, 2018.

[29] C. Gobl, "Voice source dynamics in connected speech," *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, vol 1, pp. 123-159, 1988.

[30] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication,* vol. 18, pp. 131-138, 1996.