Interspeech 2018
2-6 September 2018, Hyderabad, India

Voice Source Contribution to Prominence Perception: R_d Implementation

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

Trinity College Dublin, Ireland

murpha61@tcd.ie, yanushei@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

Abstract

This paper explores the contribution of voice source modulation to the perception of prominence, following on previous analyses of accentuation, focus and deaccentuation. A listening test was carried out on a sentence of Irish with three accented, prominent syllables (P1, P2, P3). Using inverse filtering and resynthesis, a 'flattened' version was generated, with only slight declination of f_0 and other voice source parameters. The global waveshape parameter R_d was modulated to provide (i) source boosting (tenser phonation) on either P1 or P2, and/or (ii) source attenuation (laxer phonation) following (Post-attenuation) or preceding (Pre-attenuation) P1 or P2. R_d variation was achieved in two different ways to generate two series of stimuli. f_0 was not varied in either series. Twenty-nine listeners rated the prominence level of all syllables in the utterance. Results show that the phrasal position (P1 vs. P2) makes a large difference to prominence judgements. P1 emerged as overall more prominent and more readily 'enhanced' by the source modifications. Post-attenuation was particularly important for P1, with effects equal to or greater than local P-boosting. In the case of P2, Pre-attenuation was much more important than Post-attenuation.

Index Terms: voice source modulation, prominence, R_d, prosody, perception, accentuation

1. Introduction

This paper is part of a wider project, to elaborate the role of the voice in prosody, i.e. the voice source, including not only f_0 but also other source parameters. This is desirable to gain a fuller understanding of the phonetic nature of prosody, its production and perception, as well as a more adequate account of the functions of prosody in speech communication. This has many implications in speech technology. For example, our current parallel research developing synthetic voices for Irish dialects (www.abair.ie and [1]) and deploying these in interactive games and applications for language learning (e.g., [2] and [3]) requires a more sophisticated modeling of prosody in synthesis than is currently available.

Earlier, production-based analyses of the voice source correlates of accentuation [4], focus/deaccentuation [5] and declination [6] have demonstrated prominence-lending source adjustments, often involving shifts along the tense-lax dimension of phonation. These source modulations appear to occur in synergy with the kinds of f_0 modulation which have been widely described in the literature, e.g., in [7-12]. Furthermore, the source analysis of accentuation in some Irish data [4] showed that f_0 salience is a frequent but not a necessary condition for accentuation – at least in prenuclear position. An analysis of focal accentuation also demonstrated that the adjustments entail not only the *local* source boosting of the accented syllable, but also more global effects involving the attenuation of other parts of the utterance.

An initial exploration of the perceptual impact of these types of source modulation on the perception of prominence was reported in [13]. In this study, a baseline sentence 'We were away a year ago' with 'flattened' prosody (constant settings of f_0 and other source parameters) was manipulated so that (i) one or other of the accentable syllables 'way' and 'year' was boosted (rendered with a tenser phonatory mode) and/or (ii) the material following the boosted syllable was relatively attenuated (towards a lax phonation). The results showed a striking difference in the perceived prominence of 'way' and 'year': the prominence-lending source manipulations added greatly to the perceived prominence on 'way', but relatively little to 'year'.

This leads to consideration as to whether phrasal position is important to the realization of prominence. A phrase-final (default nuclear) accent may simply require f_0 salience (e.g., a falling tone) to be prominent, where other source boosting changes alone may suffice in other positions – something suggested by the production data in [4]. Furthermore, in the above experiment, the short tail following 'year' might have reduced the potential for post-accent attenuation to have an effect. Other factors, not related to phrasal position may also have contributed, such as the difference in the vowel qualities in 'way' and 'year'.

To mitigate some of the above factors, the present study uses, as a baseline, a sentence of Irish with 3 accented syllables (referred to as P1, P2 and P3), all with the same vowel quality. The source parameters of the baseline were initially 'flattened', and a slight declination imposed to improve overall naturalness (see details in the following section). The manipulations (source boosting of the accented syllable and attenuation in preceding or following material) targeted only P1 and P2, avoiding the final accented syllable. The prediction was that the source manipulations would have roughly similar effects on their perceived prominence relative to each other.

The voice source manipulations for the individual stimuli were carried out using the global waveshape parameter R_d, proposed by Fant [14, 15] to control the tense-lax dimension of voice variation. As the objective of this experiment was to examine the role of voice source parameters (other than f_0) in prominence perception, the manipulations for the individual stimuli did not include f_0 manipulation.

2. Material and method

2.1. Baseline stimulus

The baseline stimulus was generated based on the following declarative sentence of Irish (the accented syllables are shown in bold caps):

10.21437/Interspeech.2018-2352

Bhí **CÁIT** cupla **LÁ** ar an **TRÁ**laer

[vʲi kɑtʲ kʊplʸə lʸɑ ərʸ ənʸ trʸɑlʸərʸ]

'was Kate couple days on the trawler' (word gloss)

'Kate was a couple of days on the trawler' (translation)

The sentence was produced by a male speaker of Irish (Kerry dialect) and was elicited with broad focus. The utterance was first manually inverse filtered using the interactive inverse filtering system described in [16]. Voice source parameterization was conducted using the Liljencrants-Fant (LF) model [17]. The utterance was resynthesized using cascade formant synthesis. Initially, in the baseline stimulus, all source parameters were flattened, i.e. the f_0 and R_d values were set to the averaged values across the original utterance. Following this, to improve naturalness, sentence declination effects were included, entailing a drop of 5% (3.24 Hz/s) in f_0 over the utterance, along with an increasingly more lax phonation (which corresponds here to a rise in R_d of 20% or 0.114 units/s). Our earlier analysis of declination [6] shows that phonatory quality also alters with declination, and not simply f_0 as is often assumed.

2.2. Implementation of R_d variation

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). Note that U_p/E_e is equivalent to the glottal pulse declination time T_d during the closing phase of the glottal cycle. The scale factor (0.11^{-1}) makes the numerical value of R_d equal to the pulse declination time in milliseconds when f_0 is 110 Hz [14]. Variation in R_d is proposed to reflect voice source variation along the tense-lax continuum; the values typically range between 0.5 (tense voice) to 2.5 (breathy voice).

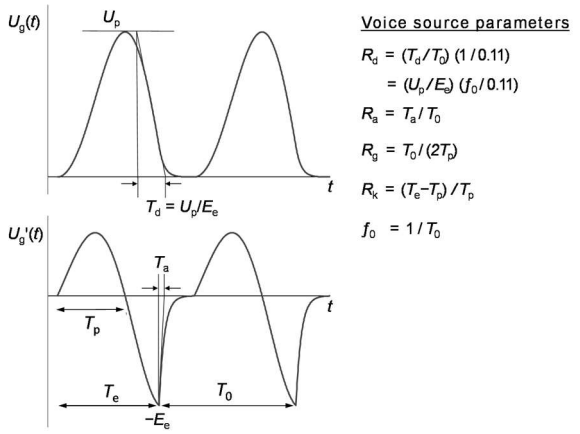

Figure 1: *Voice source parameters used to generate the LF model waveform (adapted from [14]). Upper panel: glottal flow; lower panel: glottal flow derivative.*

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. (For a description of the various glottal parameters, see [18]). 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 [14] (see also [19]).

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. Given that the intention was not to vary f_0 (other than for the sentence declination baseline as explained above), to implement the R_d variation in these stimuli, one can vary either E_e or U_p, or a combination of the two. In the earlier experiment [13], only E_e was allowed to vary. In the present study, two series were prepared. In the first, changes to R_d were implemented by varying E_e (while not modifying U_p), or by varying U_p , without modifying E_e.

In the earlier perception experiment [13], R_d was varied by controlling E_e only. In this experiment both options are tested, and two sets of stimuli (and two baseline sentences) were thus generated. The two series are henceforth referred to as the E_e stimuli and the U_p stimuli.

2.3. The individual prominence-varying stimuli

Source manipulations for each series of stimuli entailed changes to the source that should, according to our earlier descriptive studies, boost the salience (through tenser phonation) of the accented syllable (i.e. targeting P1-Cáit or P2-Lá) and/or attenuate, reduce the salience (through laxer phonation) of those portions of the utterance before (Pre) or after (Post) the syllable in question (P1-Cáit or P2-Lá).

The labels used for the individual stimuli and the specific R_d adjustments involved are glossed in Table 1 and illustrated schematically in Figure 2. Note: greater phonatory tension entails lower R_d, laxer phonation entails an increase in R_d values.

Table 1: *R_d stimulus manipulations and labels used.*

Stimulus label	R_d adjustments
Baseline	No salience-lending adjustments
Peak	Tenser phonation (lower R_d) in P1 or P2
Post	Laxer phonation (raised R_d) *after* the targeted syllable (P1 or P2)
Pre	Laxer phonation (raised R_d) *before* the targeted syllable (P1 or P2)
Peak+Post	Combinations of above adjustments
Pre+Peak	
Pre+Post	
Pre+Peak+Post	

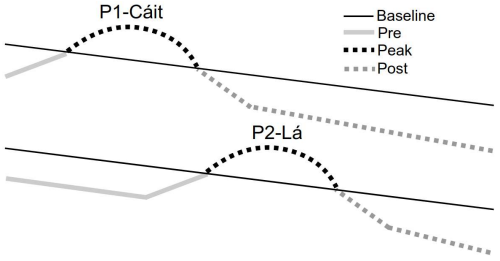

Figure 2: *Phonatory shifts from the baseline (solid black): Peak (P1 or P2) = tenser phonation (dotted black); Pre (solid grey) or Post (dotted grey) = laxer phonation.*

The R_d parameter was manipulated in three ways, using values prompted by our earlier analytic studies [5, 20, 21]. For the Peak stimuli, R_d was lowered by 33% in the targeted syllable (whether P1 or P2). In the Peak stimuli targeting P1-Cáit, the R_d value for P1 was 0.8 (compared to the P1 baseline value of 1.14). For the Peak stimuli targeting P2-Lá, the R_d value for P2 was 0.85 (compared to its baseline value of 1.27). Recall that lower R_d values correspond to tenser phonation. These values

reflect the influence of the R_d declination already applied to the baseline stimulus, as explained above.

In the Post stimuli, R_d was raised by 30% relative to the baseline value in the syllable immediately following the peak target. This was followed by a further 20% rise over the remainder of the utterance. Pre-target attenuation involved raising R_d in pre-target material by 20% relative to the baseline, with a lowering of R_d across the final pre-target syllable. The combined stimuli simply involved the combination of these adjustments.

Two sets were produced, the E_e stimuli and the U_p stimuli, with 13 stimuli in each set. Each set included, in addition to the Baseline stimulus, a range of stimuli that would potentially entail enhancement to prominence on either the P1-Cáit target or the P2-Lá target: Peak, Post, Peak+Post, Pre+Peak, Pre+Post, and Pre+Peak+Post. The Pre stimulus on its own was not included.

2.4. Listening test

The listening test was carried out with 29 participants, speakers of Irish, over headphones. In the test, the U_p stimuli were first presented in random order, followed by the E_e stimuli, also in random order. Listeners heard each stimulus four times with a two second gap between each repetition, and there was a 15 second gap before the next stimulus was presented. The participants were asked to (1) indicate the perceived prominence of each syllable in the utterance on a visual analogue scale, (2) indicate how natural the audio sounded on a scale from 1 (not natural) to 5 (natural). Each new stimulus was introduced with a beep followed by the number of the stimulus, as well as a beep 10 seconds before a new stimulus was introduced. Five random stimuli were included at the beginning of the test to allow the participants to get used to the procedure; these responses were not included in the analysis.

3. Results

The syllable prominence magnitude values obtained for each stimulus were normalized to account for the variation in the use of the scale range. Figure 3 shows the difference in perceived prominence of the two accents P1 and P2 (within the same stimulus) for the E_e stimuli (upper panel) and U_p stimuli (lower panel). Blue bars show the P1 minus P2 difference for those manipulations targeting P1-Cáit. Green bars show P2 minus P1 values for manipulations targeting P2-Lá. Thus, blue bars above zero indicate that P1 in the P1-Cáit stimuli is perceived as more prominent than P2; green bars above zero show where P2 (in the stimuli targeting P2-Lá) is deemed the more prominent. Black asterisks indicate significant differences in the magnitude of P1 and P2 within the same stimulus established by a one-way ANOVA.

Figure 3 shows the E_e stimuli (upper panel) to be much more effective in achieving prominence than the U_p stimuli (lower panel), and discussion will therefore focus on the former.

P1 is deemed more prominent than P2 in most cases: P2 is judged more prominent than P1 in only two P2-Lá targeted stimuli (E_e: Pre+Peak and Pre+Peak+Post).

It is also clear from responses that in the baseline stimulus, P1 was perceived to be more prominent than P2. Red asterisks show where the magnitude of the difference between P1 and P2 is significantly different from the magnitude of the difference between them in the baseline stimulus. If one considers the prominence difference judgements relative to the baseline, it somewhat changes the picture: one notes that the R_d manipulations have a greater and more significant effect on P2 perception – even if the baseline bias towards P1 prominence means that P2 still rarely emerges as more prominent than P1. When the baseline difference is considered, the 'enhancement' of P1 brought about by the P1-Cáit manipulations appear less dramatic. Despite this factor, P1 emerges as both intrinsically more prominent and more readily enhanced perceptually.

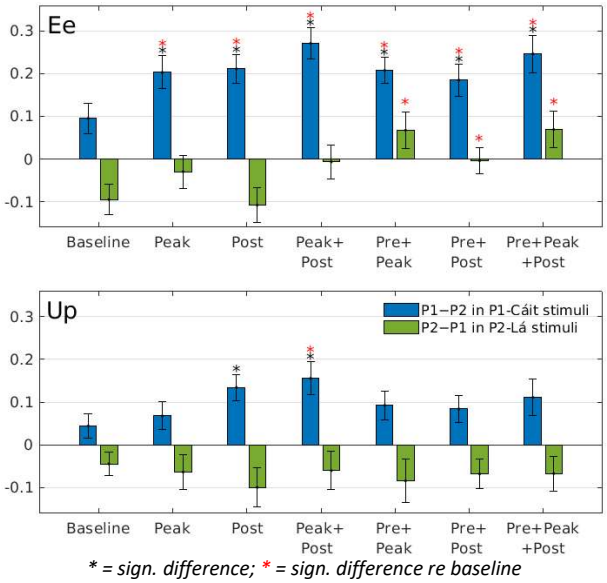

Figure 3: *P1-P2 differences for P1-Cáit stimuli (blue), and P2-P1 differences for P2-Lá stimuli (green).*

Figure 4 provides a more detailed view of the effects on the overall P1 P2 P3 contour of the P1-Cáit targeted manipulations (left panels) and the P2-Lá targeted manipulations (right panels).

Panel (a) of Figure 4 shows, relative to the baseline (black dashed line), values for those stimuli with peak-boosting, and those with post-attenuation. For P1, peak-boosting has the effect, not only of raising prominence on P1, but also of lowering it on P2. The Peak stimulus yields essentially the same contour as the Post stimulus. This suggests the functional equivalence of these two source adjustments in the utterance, the *local* and the *global*. In the case of the P2-Lá targeted stimuli, these two effects do not appear to be equivalent. Peak-boosting does raise the prominence of P2 relative to P1 and P3, but post-attenuation has little effect on the contour (though there is a drop in overall level).

Panel (b) of Figure 4 illustrates the additive effect of peak-boosting and post-attenuation. When targeting P1-Cáit, the effects are additive, increasing the perceived prominence of P1 beyond that achieved by either of the individual source adjustments. This is not so for the P2-Lá targeted stimulus: the contour is much like that of the peak-boosted P2, indicating again that post-attenuation is not contributing much.

Panel (c) compares Pre+Peak (pre-attenuation in combination with peak-boosting) and peak-boosting on its own. The effects on perceived prominence magnitude are apparent for both P1 and P2 targets. This was unexpected in the case of P1, given that the pre-attenuation for P1 pertained only to a single short unaccented syllable.

Panel (d) illustrates the combined effect of all three manipulations: Pre+Peak+Post (pre-attenuation, peak-boosting and post-attenuation). While this complex manipulation is effective for both P1 and P2, it looks like the combination of pre- and post-attenuation with peak-boosting is not necessarily more effective than the simpler combination of Peak+Post (peak-boosting and post-attenuation) for P1, and only marginally more effective than the combination of pre-attenuation and peak-boosting for P2.

The naturalness of the E_e and U_p varying stimuli was not significantly different. The average naturalness of the E_e stimuli was rated 3.2 (range 2.89-3.54), the average naturalness of the U_p stimuli was 3.1 (range 2.78-3.26).

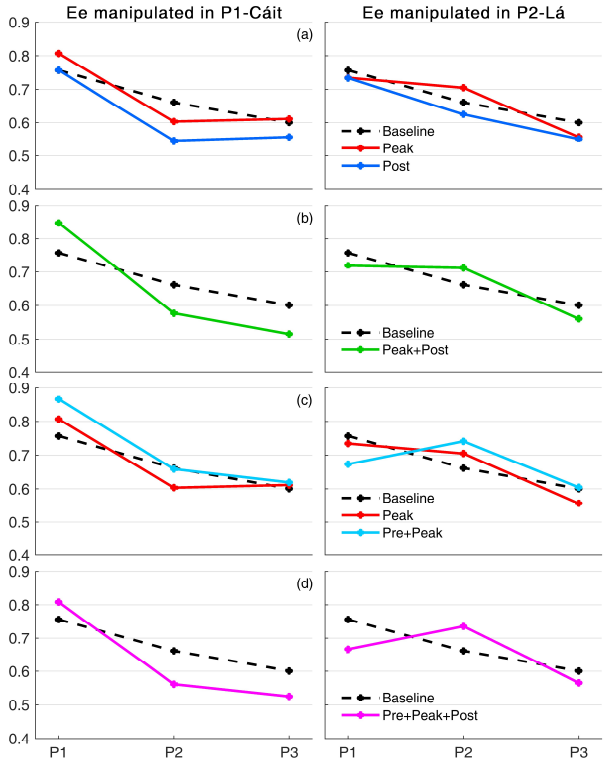

Figure 4: *P1 P2 P3 prominence for stimuli targeting P1-Cáit (left) and P2-Lá (right).*

4. Discussion

Although much is known about the contribution of f_0 to prominence (e.g., [7-12]), little is known about the contribution of other source parameters. This experiment was prompted by earlier production studies of accentuation and of focalization, which demonstrated the involvement of the entire source signal in prominence, indicating that f_0 and other source parameters work synergistically. The combined source effects entail local salience boosting to enhance prominence on the accented syllable, as well as more global, phrase-level source attenuations in other portions of the utterance. The present results illustrate how the phrase level pre- and post-attenuation may be perceptually equivalent to the peak-boosting source adjustments.

Initial expectations that the effects of source prominence manipulation on P1 and P2 would be the same were not borne out, insofar as P1 was overwhelmingly deemed the more prominent of the two. In all the P1-Cáit targeted stimuli and in all but two of the P2-Lá targeted stimuli, P1 was judged the more prominent. Furthermore, these two possible targets differed in the extent to which they were influenced by phrase-level attenuation. For P2, pre-attenuation emerged as being very effective: P2 was perceived as more prominent than P1 only in the two stimuli where pre-attenuation was present. Surprisingly, post-attenuation appeared to contribute relatively little. In the case of P1 it was very different in that both pre- and post-attenuation were effective in conferring prominence.

The dominance of P1 is very clear in the baseline values. Two possible explanations spring to mind. It may simply be that a phrase-initial accent is inherently a more prominent entity. On the other hand, it may be that the declination included in these stimuli introduced a P1 bias, conferring greater prominence in the order of P1>P2>P3. These two explanations could be complementary: an inherent prominence of the phrase-initial accent may be inextricably linked to the fact of declination. Our thinking on declination to date was that it did not in itself confer prominence, and that accentuation is judged relative to the declination line. This assumption may need to be explored further, and it may turn out that the declination-prominence 'adds' a natural prominence bias to the initial prenuclear accent of a phrase. It is worth noting, however, that in the earlier source prominence experiment [13] mentioned above, there was no declination included in the stimuli, and judgements also indicated a similar disparity in the prominence ratings of successive accents.

Of the two R_d implementations used in this experiment, the one involving variation of E_e was considerably more effective in achieving prominence than the implementation involving variation of U_p. When considering the spectral correlates of U_p and E_e variation, this finding is perhaps not too surprising. U_p is associated primarily with the level of the first harmonic, and the low end of the source spectrum, while E_e has a more pervasive influence on overall spectral levels above $H1$.

5. Conclusions

This experiment demonstrates how voice source modulations affecting phonatory quality are perceptually important in signaling prominence, operating both at a *local* level in boosting the prominence of the accented syllable and at a *global* level in attenuating the prominence of other portions of the utterance.

Results also suggest that the phrasal location may be important, revealing a bias towards greater prominence on the phrase-initial prenuclear accent, when compared to the following one. It is unclear from these data whether this bias results from the effects of voice source declination (described in [6]). This is something that will require further investigation.

One of our goals is to control for voice prosody modulation in Irish synthetic speech [1]. The global waveshape parameter R_d has been proposed as a global control parameter that can be implemented in synthesis. Two implementations of the R_d parameter were used in this study and a clear difference emerged in terms of their effectiveness in signaling prominence. A fuller elaboration of how to optimally implement this parameter will also be the focus of future work.

6. Acknowledgments

This research is supported by the Govt. of Ireland, An Roinn Cultúir, Oidhreachta agus Gaeltachta – the Department of Culture, Heritage and the Gaeltacht (*ABAIR* project).

7. 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] A. Ní Chasaide, I. Yanushevskaya, J. Kane, and C. Gobl, "The Voice Prominence Hypothesis: the interplay of F0 and voice source features in accentuation," in *Interspeech 2013*, Lyon, France, 2013, pp. 3527-3531.

[5] I. Yanushevskaya, C. Gobl, J. Kane, and A. Ní Chasaide, "An exploration of voice source correlates of focus," in *Interspeech 2010*, Makuhari, Japan, 2010, pp. 462-465.

[6] A. Ní Chasaide, I. Yanushevskaya, and C. Gobl, "Prosody of voice: declination, sentence mode and interaction with prominence," in *XVIIIth International Congress of Phonetic Sciences*, Glasgow, UK, 2015, pp. 1-5.

[7] 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.

[8] M. Vainio and J. Järvikivi, "Tonal features, intensity, and word order in the perception of prominence," *Journal of Phonetics,* vol. 34, pp. 319-342, 2006.

[9] D. J. Hermes, "Stylization of pitch contours," in *Methods in Empirical Prosody Research*, S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek*, et al.*, Eds., ed Berlin: Walter de Gruyter, 2006, pp. 29-61.

[10] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *Journal of the Acoustical Society of America,* vol. 89, pp. 1768-1776, 1991.

[11] J. Terken, "Fundamental frequency and perceived prominence of accented syllables. II. Nonfinal accents," *Journal of the Acoustical Society of America,* vol. 95, pp. 3662-3665, 1994.

[12] R.-A. Knight, "The shape of nuclear falls and their effect on the perception of pitch and prominence: peaks vs. plateaux," *Language & Speech,* vol. 51, pp. 223-244, 2008.

[13] 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.

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

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

[16] C. Gobl and A. Ní Chasaide, "Techniques for analysing the voice source," in *Coarticulation: Theory, Data and Techniques*, W. J. Hardcastle and N. Hewlett, Eds., ed Cambridge: Cambridge University Press, 1999, pp. 300-321.

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

[18] C. Gobl and A. Ní Chasaide, "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds., 2 ed Oxford: Blackwell Publishing Ltd, 2010, pp. 378-423.

[19] C. Gobl, *The Voice Source in Speech Communication. Doctoral thesis*. Stockholm: KTH, Department of Speech, Music and Hearing, 2003.

[20] A. Ní Chasaide, I. Yanushevskaya, and C. Gobl, "Voice source dynamics in intonation," in *XVIIth International Congress of Phonetic Sciences*, Hong Kong, China, 2011, pp. 1470-1473.

[21] 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.