Global waveshape parameter $R_d$ in signaling focal prominence: Perceptual salience in the absence of $f_0$ variation

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

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, 
Trinity College Dublin, Ireland

* Correspondence: 
Irena Yanushevskaya 
yanushei@tcd.ie

Word count: 11,460, 9 figures, 6 tables [+2 Supplementary figures, 1 Supplementary table]

Keywords: voice source, $R_d$, prominence, deaccentuation, perception

Abstract

This paper explores perceptual salience of voice source parameter manipulation in signaling prominence in the absence of $f_0$ variation. Synthetic stimuli were generated based on an inverse filtered all-voiced utterance ‘We were away a year ago’. A global waveshape parameter $R_d$ was manipulated in the stimuli to enhance prominence in the two potentially accentable syllables WAY and YEAR and to provide voice source deaccentuation post-focally. The manipulations were intended to mimic an increase in phonatory tension in the prominent syllable while decreasing it in the post-focal material. $f_0$ was kept constant. Two listening tests were conducted in which participants rated the perceived prominence of the potentially accentable syllables in the manipulated utterances on a continuous visual analogue scale. The results suggest that perceived focal prominence can be achieved by source variation in the absence of $f_0$ modulations, although the results were not identical in the two tests. The extent of the enhancement of prominence by source manipulations in our data depended on the location of focal syllable in the intonational phrase and on the length of postfocal material (the effect was stronger for WAY than for YEAR).

Introduction

This paper explores the perceptual salience of voice source parameter manipulations in generating prosodic prominence in synthetic stimuli when $f_0$ variation is not included.

Prosody of speech is typically described as modulations in speech melody ($f_0$/pitch variation) and timing as well as adjustments in phonatory settings (Botinis et al., 2001; Wagner & Watson, 2010; Xu, 2011), but may also involve changes in articulatory settings, e.g. Keating (2003).

The main body of work on prosody has been conducted on $f_0$, timing and intensity, both in production and perception studies, and the advances in the field in relation to these phenomena are considerable. $f_0$ and intensity are typically treated and studied as separable from the overall voice source modulations in prosody. This mainly has to do with the relative ease of measuring them...
compared to other features of the voice source perceptually correlated with voice quality, such as spectral tilt or glottal pulse shape. Intensity is often cited as a separate feature, e.g., Wagner and Watson (2010) but is to a large extent a reflection of source variation.

The role of the voice source in shaping linguistic prosody, although increasingly the focus of interest of phoneticians and speech scientists, is still understood to a much lesser extent (d’Alessandro, 2006). Voice source variation in signaling focus, deaccentuation, prominence, phrase boundaries has been explored in analytical production studies and [to a lesser extent] in perception studies, e.g., Epstein (2003); Gobl (1988); Iseli et al. (2006); Ludusan et al. (2021); Ní Chasaide et al. (2011); Pierrehumbert (1989); Strik and Boves (1992).

The main research interest here is the prosody of the voice (or voice prosody). By voice prosody we mean the modulations of the voice source in its entirety (including \( f_0 \) and other source parameters describing the shape of the glottal pulse and reflecting changes in not only what is perceived as pitch but also in voice quality and loudness) that are used to signal important linguistic information. This approach was adopted in Ní Chasaide and Gobl (2004a, 2004b).

This paper aims to investigate, using synthetic stimuli in perception tests, whether voice source modulations in the absence of \( f_0 \) variation can generate linguistic prominence. This is important for developing flexible synthetic voices where control over voice quality, and not just \( f_0 \) or intensity, is desirable while maintaining a relatively small number of control parameters, such as the global waveshape parameter \( R_d \).

1.1 Focus and deaccentuation – definition and the linguistic function of prosodic prominence

A linguistic entity is prosodically prominent if it stands out relative to its environment by virtue of its prosodic characteristics (Terken & Hermes, 2000; van Heuven, 2014; Wagner et al., 2015). Prosodic focus is generally described as a means of emphasizing, highlighting a piece of new or contrastive information relative to the information already shared by the conversation participants. Focal prominence can be signaled by a variety of phonetic and phonological properties such as the type and alignment of pitch accents, boundary tones, duration, intensity and \( f_0 \) range (Baumann & Winter, 2018; Burdin et al., 2015). These cues work synergistically to provide ‘robust communication of prominence information’ (Baumann & Winter, 2018). New information can be made salient not only through prosodic highlighting but also using various syntactic and semantic means (Kember et al., 2019). Prosodic highlighting is achieved by simultaneous de-highlighting or de-accentuation of the known, given information (Ladd, 2008).

1.2 The form/correlates of prosodic prominence – evidence from production studies

\( f_0 \) variation has been widely described as a primary acoustic correlate of focus (by assigning pitch accents to the syllables that are lexically stressed or by extending the range of \( f_0 \)), e.g., Ladd (2008), Féry (2017). For example, in English, a narrow focus is generally characterized by a high falling (H*L) nuclear pitch accent. This is accompanied by an increase in the pitch range in the focally accented syllable and compression of pitch immediately following the focused syllable and deaccentuation, particularly of the post-focal material (no change is usually observed in the pitch range of pre-focus material) (Xu & Xu, 2005). Focally accented syllables have also been found to have longer duration and higher intensity (Leemann et al., 2016; Turk & Sawusch, 1996). Languages differ in the use of these acoustic cues to signal prominence and deaccentuation (Cruttenden, 2011; Leemann et al., 2016).
Most inferences on voice source correlates of more general phenomena of stress, accent and prominence (rather than focus) have been based on spectral measurements derived from the speech waveform, which should reflect source behavior. For example, the amplitude level of H1 can be compared to the level of some higher frequency component such as H2 or A1 (the amplitude level of the first formant F1) as a measure of overall shape and slope of the spectrum. For a discussion of such measures, and more generally of how source parameters relate to the spectral characteristics of speech, see Gobl and Ni Chasaide (2010).

The spectrum-based source measures in the analyses in Sluijter and van Heuven (1996), Sluijter et al. (1997) suggested that a less steep spectral slope (boosting of the higher frequency regions) and consequently tenser voice quality is associated with focal stress and prominence in Dutch and American English. Similar findings are reported in Heldner (2003) who points to the overall intensity and spectral emphasis as reliable acoustic cues of focal accents in Swedish (spectral emphasis is a measure of a relative contribution of high frequency components to overall intensity). Spectral tilt has been reported as a reliable cue to prominence in dialogue speech in (Campbell, 1995; Campbell & Beckman, 1997).

Shue et al. (2007) compared voice source correlates of pitch accent and lexical stress. They found that stressed syllables have lower Open Quotient (indicative of tenser voice) irrespective of pitch accent, and also longer duration. Acoustic cues of lexical stress were found to be affected by the presence of pitch accent, boundary tone and by speaker gender.

More recently, Baumann and Winter (2018) used measures of spectral tilt H1-A2 and H1-A3 (difference between the amplitude level of the first harmonic and amplitude peaks near F2 and F3), along with other acoustic measures, in perception studies on word prominence in German and found that both measures of spectral slope were correlated to prominence judgment in statistically significant way.

Kakouros et al. (2018) provided a comprehensive review of spectral tilt as a correlate of prosodic prominence and explored its importance in signaling sentence prominence in Dutch and French relative to the more established acoustic correlates $f_0$, intensity and duration. They point out that measures of spectral tilt are diverse and the standards are less established than for measures of $f_0$ and intensity, e.g., some spectral tilt measures are directly computed using the speech pressure waveform (Campbell, 1995; Eriksson et al., 2001; Heldner, 2001; Sluijter et al., 1997) and others calculate spectral tilt of estimated glottal source obtained by inverse filtering (Iseli et al., 2006; Kreiman et al., 2007). The results of classification experiments on clean and corrupted speech in (Kakouros et al., 2018) suggest that measures of spectral tilt are important contributors in differentiating prominent and non-prominent words.

Analyses of source parameters (described in more detail in section 2.1.2 below) obtained by glottal inverse filtering and parameterisation generally point to changes in the shape of the glottal pulse in accented or prominent syllables that suggest greater tension in the mode of phonation (d'Alessandro, 2006; Epstein, 2002; Iseli et al., 2006; Koreman, 1995). Increased vocal effort associated with focal prominence entails greater volume-velocity airflow through the glottis, more asymmetrical glottal pulses with smaller open quotient and steeper, more abrupt glottal closure, which in turn generates relatively stronger higher harmonics and flatter spectral tilt (van Heuven, 2014).

Measures of source correlates of focal accent in Swedish using a manual interactive technique (Gobl, 1988) identified dynamic changes in the strength of the glottal excitation ($E_c$) in focal context,
enhancing the vowel-consonant distinction. These data further suggest that the focal patterning of an utterance does not just affect the focally accented syllable but may also have consequences for the pre-focal and post-focal material.

Results of studies of focus and stress in Finnish (Airas et al., 2007; Vainio et al., 2010) run counter to the general trends mentioned above. In both studies the analysis was carried out using an automatic method - Iterative Adaptive Inverse Filtering (Alku, 1992), and the data analysed were vowel segments extracted from connected speech. These studies report higher NAQ values in focally accented syllables (Vainio et al., 2010), as in stressed syllables (Airas et al., 2007). The NAQ measure (Alku et al., 2002) has been proposed as a global parameter, which correlates with the tense/lax dimension of vocal quality and, when scaled by 0.11, is essentially the same as the $R_d$ parameter used in this study. A high NAQ value is indicative of lax voice, and a low NAQ value indicative of pressed or tense voice. NAQ has gained considerable popularity as a measure of the tense/lax dimension of voice variation.

Yanushevskaya et al. (2010) found that focally accented syllables involve increased respiratory effort, with stronger excitation, a smaller Open Quotient (features essentially associated with tenser phonation). Similarly, a number of source parameters contributed to deaccentuation in the postfocal material of the utterance: falling normalized glottal frequency $R_g$, peak glottal flow $U_p$, glottal excitation $E_e$ and rising open quotient $O_q$. Ní Chasaide et al. (2015) found that declination – a downward trend of $f_0$ over the course of an utterance is realized not only in $f_0$ but also in other parameters of the source: e.g., there is declination in $E_e$, $R_g$ and $C_q$ (closed quotient, defined as $1-O_q$) indicating a reducing level in the voice source excitation strength and increasing relative dominance of the lower end of the source spectrum (increasingly lax mode of phonation). Ní Chasaide et al. (2013) sought support for the Voice Prominence Hypothesis suggesting that prominence arises from the contribution of different source parameters, and that the extent to which a particular parameter contributes can vary. Thus, for example, accentuation of syllables which have no pitch prominence is signaled by other parameters of the source. Furthermore, speakers may use different strategies resulting in different combinations of source parameters when signaling prominence (Yanushevskaya et al., 2017).

The picture emerging is that prosody entails the modulation of the entire voice source (including $f_0$) and that the different parameters appear to work synergistically in contributing to the realization of prominence, deaccentuation, etc. The findings in Ní Chasaide et al. (2013) suggested that even in the absence of $f_0$ salience, other voice source parameters appear to take over in signaling prominence.

It can be noted that, although $f_0$ and source parameters often covary, they can also be controlled independently of each other.

### 1.3 Source correlates of focus and prominence – evidence from perception studies

As pointed out by van Heuven (2014), a reliable acoustic correlate is not necessarily an important perceptual cue to prominence: relatively small changes in terms of production in $f_0$ might be highly perceptually salient. On the other hand, intensity, a highly reliable cue acoustically, may not emerge as a salient cue of prominence (van Heuven, 2014).
Although ‘it is currently still unclear which linguistic variables have the strongest impact on the perception of prominence’ (Baumann & Winter, 2018), in general, the existing body of research suggests that $f_0$ and to a somewhat lesser extent duration are of greatest perceptual importance in signalling prominence (Gordon & Roettger, 2017; Kakouros et al., 2018; van Heuven, 2014; Wagner et al., 2015).

A study into relative cueing power of $f_0$ and duration in German (Niebuhr and Winkler, 2017) found that ‘an increase in $f_0$ of less than 1 semitone is needed in order to outweigh an increase in duration of 30% on a neighbouring syllable.’ Baumann and Winter (2018) studied how acoustic parameters, discrete prosodic categories and non-prosodic (e.g., semantic, syntactic) factors interact to signal prominence, and used a random forest classification algorithm to establish which of them were of relatively higher salience in a prominence identification task. While their findings support the general view that multiple cues to prominence interact, pitch accent position and type emerged as the most salient cue for prominence.

Intensity is generally viewed as a relatively minor cue to perceived prominence, stress/focus, e.g., Fant and Kruckenber (1994). According to van Heuven (2014), intensity has a perceptual effect only if its increase or decrease is concentrated in frequency bands above 500 Hz, thus affecting spectral slope.

The above views are not universally accepted. Some studies suggested that duration (Heldner, 1998) and loudness (Kochanski et al., 2005) may be more important cues to signaling prominence than $f_0$. As pointed out in a number of studies, cues operate synergistically. For example, Kuang and Liberman (2018) demonstrated that voice quality cues (spectral slope) are used by the listeners as an indicator of pitch range and affect their perception of pitch height. In a recent paper, Ludusan et al. (2021) explored the impact of CPP - Cepstral Peak Prominence (Hillenbrand et al., 1994), which they used as a measure of voice quality, alone and in combination with $f_0$, duration and intensity, on perceived syllable prominence ratings by naïve and expert listeners in German. They report ‘stable but subtle’ effect of voice quality cues on prominence perception: CPP cues are used by the listeners to identify prominent syllables and have a significant effect on prominence ratings. Random forest analysis showed, however, that duration and intensity (RMS) cues appeared more important than voice quality (CPP) and $f_0$ for both expert and naïve participants.

While there have been many experimental studies demonstrating the role of $f_0$ peaks in signaling prominence, accentuation and focus (Gussenhoven et al., 1997; Hermes, 2006; Knight, 2008; Kuang & Liberman, 2018; Pierrehumbert, 1979; Terken, 1991, 1994; Vainio & Järvikivi, 2006), there is little on the perceptual role of voice source adjustments other than $f_0$. Fant and Kruckenber (1996) wrote: ‘We are beginning to understand most of the basic phenomena but we lack systematic and sufficiently complete descriptions. A problem is that we have very little experience from perceptual experiments. Much work is needed to reach an insight in the relative perceptual salience of various components of a source rule system.’ (p.45). This paper, we hope, will be among those to provide such an insight.

1.4 Voice prosody and prominence manipulation in synthesis

It has been shown that voice quality is an integral part of linguistic and paralinguistic prosodic signaling. Natural and intelligible speech synthesis must use ‘correct’ prosody and must include voice quality variation (d'Alessandro, 2006).

Including voice quality features is desirable if the goal is to develop natural expressive speech synthesis. Voice quality manipulations have been implemented into speech synthesis with the goal of producing more expressive synthetic voices in a number of applications.
In d’Alessandro and Doval (2003), speech units within a concatenative speech synthesizer could be modified by manipulating the magnitude spectrum of the periodic source components, therefore changing the spectral tilt and glottal formant of the source. The method of expressive synthesis in Cabral and Oliveira (2006) uses pitch-synchronous time-scaling to modify the LPC residual of speech in order to transform \( f_0 \) and other source parameters related to voice quality. Neutral speech samples were resynthesized using global transformations of voice source parameters derived from emotional speech data of several different affective states. Another method, developed by Cabral et al. (2011), allows for the control and transformation of the voice source in statistical parametric speech synthesis (SPSS) by removing the effects of the source from speech, and then replacing it with a synthetic LF-model based source signal. This parameterized signal can be transformed to change the voice quality. This method still requires a level of expert knowledge to carry out any transformations as it does not include a control interface. In addition, the synthetic speech produced by this system is sensitive to the errors in the analysis of the source. Control of voice quality along the tense-lax dimension in concatenative speech synthesis in Buchanan et al. (2018) was achieved by using LF-model pulses in place of the voice source. However, the transformations lead to a substantial drop in perceived naturalness.

Our earlier production and perception studies (Gobl et al., 2002; Gobl & Ní Chasaide, 2003b; Ryan et al., 2003; Yanushevskaya et al., 2018) involved detailed analysis and synthesis of many voice source parameters, many of which covary in natural speech production. Controlling such an array of parameters in synthesis would be very difficult, and would render the system unusable by all but experts in the field. One approach that has been used in speech synthesis applications is to reduce the number of parameters and to explore the use of a global waveshape parameter – the \( R_d \) parameter (Fant, 1995, 1997), also described in detail in section 2.1.3 – to control voice quality modulations in synthetic speech.

Degottex et al. (2013) describe how the \( R_d \) parameter can be successfully used in the modification of breathiness in a HMM-based synthetic voice, while Huber and Roebel (2015) used variations in \( R_d \) to produce voice qualities from very tense to very lax. Sorin et al. (2017) used \( R_d \) manipulations as a means to transform the voice quality of synthetic speech by adjusting the source signal of voiced speech in a concatenative speech synthesis system, so that it became, what they called, semi parametric.

### 1.5 The aims of the current study

Exploring the use of the \( R_d \) parameter to control for the tense-lax dimension of voice quality in prosodic variation through perception experiments sets out to perceptually test the production findings, reviewed in the previous section, concerning the tense-lax source modulations associated with focus, accentuation and deaccentuation. It would also contribute to the development of flexible prosodically rich speech synthesis systems. It should be noted that in this study the voice quality alone is manipulated. This is done in order to demonstrate the importance of voice quality dimension of prosody – even in the absence of the \( f_0 \) cues which typically accompany voice quality modulation in real speech.

Most models of prosody do not account for any voice source parameters other than \( f_0 \). In exploring the perceptual contribution of voice quality dimensions (here the \( R_d \) parameter) to focal prominence, this study hopes to contribute to a deeper understanding of how \( f_0 \) and other source parameters combine/covary, which is important in developing prosodic models accounting for voice quality variation.
This paper explores the perceptual importance of voice source adjustments which we have observed in sentences with variable location of focal accent. We further aim to elucidate whether such voice source adjustments on their own, without \( f_0 \) salience, might be capable of shifting the perception of the location of focal accent within the sentence.

In this study a recording of the all-voiced sentence ‘We were away a year ago’, produced with broad focus, was analyzed and subsequently manipulated so that the two accentable syllables WAY and YEAR were subjectively deemed to have the same degree of prominence. This served as the baseline stimulus. Voice source characteristics were then further manipulated in ways that should in principle enhance the prominence of one or other of these syllables. Stimuli were constructed in which the voice source was manipulated in the potentially accentable syllables WAY and YEAR as well as in the postfocal part of the utterance.

The main research questions are:

(1) To what extent can such source manipulations (here \( R_d \)) induce the perception of focal accent on one or other syllable?

(2) Which of the source manipulations (or which combinations of source manipulations) are most effective in cueing focal accentuation?

In the experiments reported here, \( f_0 \) did not vary across the stimulus set. This is not to suggest that \( f_0 \) does not play a major role in cueing focus but rather an attempt to explore how voice factors other than \( f_0 \) might be contributing, and to see whether source variations alone (without \( f_0 \) variation) can alter the perception of where the focal accent lies in a phrase. The extent of source variation used in these stimuli falls well within the ranges observed in production studies.

2 Material and Method

2.1 Synthetic stimuli

2.1.1 Speech material

The stimuli were constructed on the basis of an all-voiced utterance ‘We were away a year ago’ produced by a male speaker of Irish English. The utterance was elicited with broad focus (with the focal prominence realized by the speaker on YEAR) and was recorded as part of another study (Gobl et al., 2015), where further versions of the sentence with a focal accent on the syllables WAY and YEAR were also obtained and source characteristics analyzed. The utterance was manually inverse filtered using interactive inverse filtering software (Gobl & Ní Chasaide, 1999; Ní Chasaide et al., 1992). Voice source parameterization was subsequently conducted using the Liljencrants-Fant (LF) model (Fant et al., 1985).

2.1.2 Important voice source parameters

Important characteristics of the voice source pulse shape can be captured by source parameters, such as \( E_e \), \( R_a \), \( R_k \), \( R_g \), \( O_q \) and \( U_p \) (see, e.g., Gobl & Ní Chasaide 2010). \( E_e \), the excitation strength, is the negative amplitude of the differentiated glottal waveform at the time point of maximum change in the waveform derivative. It relates to the overall strength of the glottal excitation. \( R_a \) is the normalized effective duration of the return phase, i.e. the interval for which the glottis remains open after the main excitation. \( R_a \) relates to the spectral slope: the higher the \( R_a \) value, the greater the spectral slope. \( R_k \) is a measure of glottal pulse symmetry, defined as the duration of the closing portion of the pulse
Relative to the duration of the opening portion. Thus, a lower $R_k$ value means a more skewed pulse.

$R_g$, the normalized glottal frequency, is a measure of the characteristic frequency of the glottal pulse ($F_g$), normalized to $f_0$. $R_g$ mainly affects the relative amplitudes of the low end of the source spectrum. $O_q$, the open quotient, is a measure of the open phase of the glottal pulse as a proportion of the glottal period. $O_q$ can be determined entirely by $R_g$ and $R_k$ according to $O_q = (1 + R_k)/(2R_g)$. It thus excludes the return phase (captured instead by the $R_a$ parameter). $O_q$ mainly affects the amplitudes of the lower end of the source spectrum. $U_p$, the peak glottal flow, is a measure of the maximum amplitude of the glottal flow pulse (Gobl et al., 2019), see also Figure 1.

The source parameters tend to covary, and the global waveshape parameter $R_d$, described below, aims to take this covariation into account. It is a parameter that defines the overall shape of the LF-model pulse waveform capturing some of its important characteristics in one single measure. As a long-term aim is to explore the possibilities to control prosody in speech synthesis by using a limited set of parameters, the current work is focused mainly on $R_d$.

### 2.1.3 The $R_d$ parameter

The $R_d$ parameter was proposed in Fant (1995, 1997) as an extension of the LF model and ‘a data reduction scheme whereby the waveshape parameters $R_k$, $R_g$ and $R_a$ are collapsed into a single parameter $R_d$’ (Fant & Kruckenberg, 1996, p. 47).

The $R_d$ parameter is derived from $f_0$, $E_e$ and $U_p$ as follows:

$$R_d = \left(\frac{1}{0.11}\right) \left( f_0 \cdot \frac{U_p}{E_e} \right)$$  \hspace{1cm} (1)

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 (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 (Figure 1). The scale factor ($0.11^{-1}$) makes the numerical value of $R_d$ equal to the declination time in milliseconds when $f_0$ is 110 Hz (Fant, 1995).

Variation in $R_d$ tends to reflect voice source variation along the tense-lax continuum; the values typically range between 0.3 (tense voice) to 2.7 (lax voice). Lower $R_d$ values can be used to generate tenser voice quality (flatter spectral slope, stronger higher frequency harmonics). High values of $R_d$ would result in a laxer voice and a steeper spectral slope.

Generally speaking, to synthesize the LF model glottal waveform, data for $R_d$, $E_e$, $R_a$, $R_k$ and $f_0$ parameters are required. An advantage of $R_d$ is that other parameters of the glottal source can be predicted from $R_d$ using formulas derived from linear regression analysis (Fant, 1995), e.g.:

$$R_a = \frac{-1 + 4.8R_d}{100}$$  \hspace{1cm} (2)

$$R_k = \frac{22.4 + 11.8R_d}{100}$$  \hspace{1cm} (3)
A principal component analysis carried out on various voice source parameters in Yanushevskaya et al. (2017) suggested that \( R_d \) was important in describing cross-speaker differences in the source correlates of focus. As our earlier analyses of the speaker used here suggest shifts towards tenser phonation in focally accented syllables and towards laxer phonation in the post-focal material (Yanushevskaya et al., 2010; Yanushevskaya, Ní Chasaide, et al., 2016; Yanushevskaya et al., 2017), the adjustments made to mimic these effects in our synthetic stimuli involved lowering the values of \( R_d \) in the potentially accentable syllables and raising it in the post-focal part of the utterance.

In the current study, \( R_d \) was used as a control parameter in synthesis and it is the \( R_d \) contours that were manipulated to generate synthetic stimuli. Based on the controlled changes in \( R_d \), \( E_e \) was recalculated, while \( f_0 \) and \( U_p \) were kept constant (see Section 2.1.5).

### 2.1.4 Baseline stimulus

The \( f_0 \) contour of the original broad focus utterance contained focal prominence on YEAR. Flattening of \( f_0 \) and creating a sentence with equally non-prominent/ambiguously prominent syllables seemed appropriate given that the aim of the study was to test the ability of the voice source adjustments in the tense-lax continuum to signal focal prominence. As our earlier analytic studies showed \( R_d \) decrease in focal syllables and increase in the postfocal material, \( R_d \) contour was also flattened to remove these potential cues to prominence from the baseline stimulus.

In the baseline stimulus, the values of \( f_0 \), \( R_d \) and \( E_e \) obtained in the inverse filtering and source parameterization analysis (Section 2.1.1) were first set to the global average values across the utterance (\( f_0 = 120 \) Hz, \( R_d = 0.86 \), \( E_e = 69.8 \) dB).

As the overall impression of this stimulus was that it sounded rather tense, the values of \( f_0 \) and \( R_d \) were adjusted to make it less tense and improve the naturalness: \( f_0 \) was increased by 5\% to 127 Hz and \( R_d \) was increased by 50\% to 1.3. These changes also resulted in a lowering of \( E_e \) to 67.2 dB. This version of the utterance served as the baseline for further manipulations. There was only a minor difference in syllable durations: WAY 188 ms, YEAR 180 ms.

### 2.1.5 Stimuli with source manipulations to enhance focal prominence

Synthetic stimuli were constructed by manipulating the baseline stimulus as follows. \( R_d \), \( U_p \), \( f_0 \), formant parameters (frequencies and bandwidths) and timing values were loaded in from the text file of the manual inverse filtering data. As mentioned above, \( R_d \) and \( f_0 \) means were calculated and adjusted (1.05 \( \times \) \( f_0 \) mean; 1.5 \( \times \) \( R_d \) mean). The magnitude and timing of \( R_d \) ‘peaks’ (located at the midpoint of the vowels in the syllables WAY and YEAR) in the baseline stimulus were modified and \( R_d \) in the post-focal material was adjusted in various combinations (described below). \( R_d \) contours were generated after manipulating the baseline stimulus and the \( f_0 \) mean and \( U_p \) mean values were used to calculate new \( E_e \) contours using:

\[
E_e = \frac{U_p f_0}{0.11 R_d}
\]

\( R_d \) and \( R_k \) values were predicted from \( R_d \) according to equations (2) and (3) respectively. However, since \( R_d \) is not actually a parameter of the LF model, we also needed to predict \( R_g \). In this case we used the approximate formulas for the relationship between the two parameters derived from the
amplitude-based expressions for $R_k$, $R_g$ and $E_i$ (the maximum positive value of the LF model pulse) in Gobl and Ni Chasaide (2003a).1

$$E_i = \frac{\pi \cdot R_k \cdot E_e}{2}$$ (5)

$$R_g = \frac{E_i}{f_0 \cdot u_p \cdot \pi}$$ (6)

The obtained parameter values were then used to generate LF pulses, which were concatenated and filtered using the corresponding formant frequencies and bandwidths.

Increased phonatory tension tends to correspond to a drop in $R_d$; however, for the purpose of this paper we refer to such $R_d$ drops as ‘peaks’, as it seems intuitively easier for a reader to associate increased phonatory tension with positive values. Note also that we will refer to the manipulated utterances in the text as WAY-manipulated and YEAR-manipulated ones, although the source manipulations concerned not just the potentially accentable syllables but the following material as well.

The ranges of values used in the manipulations were based on the voice analysis of the speaker in earlier production studies mentioned above (Ni Chasaide et al., 2011; Yanushevskaya et al., 2010; Yanushevskaya, Ni Chasaide, et al., 2016). $f_0$ values were kept constant in all the syllables of the stimuli. The manipulations are described below and illustrated schematically in Figure 2.

**Peak height (magnitude) in focal syllables**: Three levels of peak magnitude were used: no peak (=Baseline), low peak and high peak. The $R_d$ values were set as follows: no peak, $R_d = 1.3$; low peak, $R_d = 1.1$; high peak, $R_d = 0.9$. These changes in $R_d$ resulted in the following $E_e$ values: no peak, $E_e = 67.2$ dB; low peak, $E_e = 68.6$ dB; high peak, $E_e = 70.3$ dB.

**Peak timing**: Stimuli were also generated where peak timing was changed relative to the vowel mid-points in the syllables WAY and YEAR. Two peak timing settings were used, in addition to the default baseline value: early peak and late peak. The values were shifted by 20% relative to the duration of the vowel. Early peak corresponds to faster increase towards the peak value and slower decrease of parameter values within the syllable; later peak corresponds to a slower rate of change of parameter values to the peak and a faster decrease of the values after the peak (Figure 2). These manipulations were added, as earlier studies of focal accentuation (Gobl, 1988; Ni Chasaide et al., 2011; Yanushevskaya et al., 2010) have suggested that source dynamics are heightened at the edge of the focally accented syllable.

**Source deaccentuation in postfocal material**: Three levels of deaccentuation in the postfocal material were used: no deaccentuation (=Baseline), shallow deaccentuation and steep deaccentuation. For the WAY-manipulated sentences, deaccentuation pertains to the entire sequence ‘a year ago’,

---

1 Since these formulas provide an approximation of the relationship between the parameters of the LF pulse, some small fluctuations in Up might in fact be present in the stimuli, but these were considered sufficiently small to be disregarded.
whereas for the sentence where YEAR is manipulated, deaccentuation is necessarily limited to the syllables of ‘ago’.

The $R_d$ values were as follows: no deaccentuation, final $R_d$ value = 1.3; shallow deaccentuation, final $R_d$ value = 1.52 (equivalent to a 20% increase in $R_d$); steep deaccentuation, final $R_d$ value = 1.78 (equivalent to a 40% increase in $R_d$). These adjustments would result in two different rates of change for the WAY and YEAR stimuli due to the different duration of post-focal material. Thus the rate of change was for WAY shallow 0.5 units/s, for WAY steep 1 unit/s, for YEAR shallow 1.35 units/s, for YEAR steep 2.7 units/s. These changes in $R_d$ resulted in following changes is $E_e$: no deaccentuation, final $E_e$ value = 67.2 dB; shallow deaccentuation, final $E_e$ value = 65.6 dB; steep deaccentuation, final $E_e$ value = 64.3 dB.

Stimuli were generated in which peak magnitude, peak timing and deaccentuation were manipulated individually and in combinations. The combinations of manipulation types are shown in Table 1. Overall, 20 combinations were synthesized for each of the WAY-manipulated and YEAR-manipulated sentences. The total number of stimuli used in the listening tests was 41 (2 syllables × 20 combinations + 1 baseline stimulus). It should be noted that differences in peak magnitude and peak timing correspond to the rate of signal change. The extent of manipulation is within the ranges found in natural human speech production. Informal auditory analysis by the authors suggested adequate quality of the utterances and audible shifts in prominence in the manipulated utterances.

[Insert Table 1 about here]

2.2 Listening tests

Two separate online listening tests were conducted in which the 41 synthesized stimuli were presented to participants in random order. In both tests, the participants were advised to use high quality headphones during the tests. The participants were informed that they were going to hear utterances in which the syllables WAY or YEAR may or may not be realized as prominent. They were asked to listen to each stimulus as many times as they wish and to complete several tasks.

2.2.1 Test 1 tasks

In the first test, the participants’ tasks were as follows:

1) Select the prominent syllable (‘Which word is the most prominent?’ WAY, YEAR, Neither);
2) For the prominent syllable, indicate the magnitude of prominence, using a slider on a continuous visual analogue scale;
3) Indicate how confident you are (on a continuous visual analogue scale, ‘not at all confident – very confident’);
4) Indicate how natural the utterance sounds (on a continuous visual analogue scale, ‘not at all natural – very natural’).

2.2.2 Test 2 tasks

In the second listening test, the participants were asked to mark the relative prominence of all the syllables in the utterance by adjusting sliders on a continuous visual analogue scale, similar to the setup in Eriksson et al. (2001).
As in test 1, the participants were asked to rate the naturalness of the stimuli and to indicate how confident they were in their judgment on a continuous visual analogue scale.

The first experiment was completed by 29 participants; the second experiment was done by 28 participants. Each test took approximately 20 minutes to complete. The experimental interface for both tests is illustrated in Figure 3.

The tests were different in terms of the complexity of the task. Test 1 required a simple choice between the two syllables to indicate a more prominent one and the extent of its prominence. Test 2 required that participants assess prominence relative to the overall utterance and paying attention to the adjacent syllables. While our study examines the effect of source manipulations on signaling focal accentuation, we asked our participants to indicate perceived prominence of the syllables rather than presenting them with mini-dialogue scenarios where a narrow focus could be elicited. This decision was made to simplify the task for the participants and reduce potential listener fatigue.

2.3 Expectations

Our expectation was that the WAY and YEAR syllables in the sentences where the voice source for those syllables and for the following material was systematically manipulated would tend to be identified as more prominent relative to the baseline. We also hypothesized that the degree of prominence perceived on the targeted syllable would correlate with the magnitude of the source manipulation carried out, e.g., the higher the magnitude of Rd peak (= the lower the Rd and the tenser the voice) the more prominent the syllable would be rated. Similarly, steeper deaccentuation was expected to contribute more than the more shallow kind. Furthermore, it was expected that the combined extreme manipulations (e.g., high peak combined with steep deaccentuation) would produce higher prominence magnitude ratings compared to individual manipulations.

3 Results

3.1 Listening test 1

3.1.1 Analysis of response count data

Table 2 shows the summary confusion matrix of perception of the stimuli in Test 1. The results show a clear difference in how the prominence of WAY-manipulated and the YEAR-manipulated stimuli was rated. Overall, the WAY-manipulated sentences (the sentences in which the WAY syllable and following material were manipulated) were identified as having prominence on WAY in most cases (64%). For the YEAR-manipulated sentences, listeners were as likely to hear prominence on WAY (37%) as on YEAR (39%). The Baseline stimulus was intended to have no prominence on either WAY or YEAR; the results of the listening test appear to suggest that that was indeed the case.

The results of Test 1 for the individual stimuli are shown in Figure 4. As clear from Figure 4, there is a bias towards WAY: more stimuli were selected as having a prominent WAY syllable in the WAY-
 manipulated sentences than YEAR in the YEAR-manipulated sentences. The stimuli for which 70% or more of the listeners identified WAY as prominent were mainly those with high $R_d$ ‘peaks’ (= tenser voice) and steep postfocal deaccentuation (= greater increase in voice laxness). Interestingly, stimuli LP+Steep and Steep were frequently selected as prominent in the way-set and thus were quite effective in prominence cueing. This points at importance of postfocal deaccentuation in signaling prominence. Conversely, the stimuli which include manipulations involving a low peak or shallow deaccentuation were identified as prominent by fewer participants.

For the YEAR-manipulated stimuli, results were very different: here, there were only relatively minor shifts from the results obtained by the baseline stimulus. In the YEAR-manipulated set, none of the stimuli was selected as having prominence on YEAR by 70% or more of the participants.

Multinomial logistic regression analysis was performed to explore the relationship between the types of manipulation (peak magnitude, peak timing, deaccentuation) and the likelihood of choice of the manipulated syllable as prominent. The analysis was conducted in R (R Core Team, 2019) using mlogit package (Train, 2009). The analysis was conducted separately for the WAY-manipulated and the YEAR-manipulated stimulus sets. The reference category was ‘neither’. The parameter estimates are shown in Table 3(a, b).

**Way-manipulated stimuli.** Addition of peak magnitude (Peak), peak timing (Timing) and deaccentuation (Deac) as main predictors to a model that contained only the intercept significantly improved the fit between model and data, $\chi^2 (6) = 56.01$, McFadden $R^2 = .05$, $p < .001$. Significant unique contributions were made by Peak, $\chi^2 (2) = 27.64$, $p < .001$, and Deac $\chi^2 (2) = 35.63$, $p < .001$, but not Timing $\chi^2 (2) = 0.04$, $p = .982$. As the peak height and the steepness of deaccentuation increase, the odds of the WAY syllable being selected as prominent (relative to ‘neither’) also increase (multiplicatively by 1.93 and 1.84 respectively).

**Year-manipulated stimuli.** Similar to the above, including peak magnitude (Peak), peak timing (Timing), deaccentuation (Deac) as main predictors in the model compared to the intercept only model significantly improved the model fit, $\chi^2 (6) = 25.72$, McFadden $R^2 = .02$, $p < .001$. Significant unique contributions were made only by Peak, $\chi^2 (2) = 17.58$, $p < .001$. The parameter estimates are shown in Table 4. The reference category was ‘neither’. As the peak magnitude increases, the YEAR syllable is increasingly more likely to be selected as prominent, multiplicatively by 1.85. Deaccentuation did not emerge as an important factor, most likely because of the difference in duration of postfocal material in the WAY-manipulated and YEAR-manipulated stimuli: four syllables vs. two syllables respectively.

[Insert Table 3 about here]

### 3.1.2 Magnitude of perceived prominence

Figure 5 shows the mean magnitude of perceived prominence of WAY- and YEAR-manipulated syllables across different stimuli. Only cases where manipulated WAY and YEAR were identified as prominent by the participants in the listening test are shown. At a glance, the magnitude mean values

---

2 The model for the analysis of WAY-manipulated was: mlogit(formula = Sel ~ 1 | Peak + Deac + Timing, data, reflevel = "neither", method = "nr"); the model for YEAR-manipulated included the same predictors.
of perceived syllable prominence range between about 40 and 70 [=no extreme values]; perceived prominence magnitude is higher for the way-set than for the year-set overall, Figure 5(panel B).

To test if the type of parameter manipulation has an effect on the magnitude of perceived prominence of the manipulated syllable, a series of linear mixed effect model analyses were conducted. The cases when neither syllable was selected in the listening test as prominent were excluded from the analysis.

Analyses were conducted in the R environment (R Core Team, 2019) using the lme4 [version 1.1-29] package (Bates et al., 2015) for model fitting. The lmerTest package (Kuznetsova et al., 2017) 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 fit by maximum likelihood (ML) method. The initial model included Peak, Timing, Deaccentuation (Deac) and Target (=manipulated syllable) as the main predictor variables (fixed effects) as well as their interaction; random effects included by-subject random intercepts:

\[ \text{Magnitude} \sim \text{Target} \times \text{Peak} \times \text{Timing} \times \text{Deac} + (1 | \text{Participant}) \].

The final reduced model included Target, Peak and Deaccentuation as the fixed predictors, the Target:Peak and Target:Deac interactions and by-subject random intercepts:

\[ \text{Magnitude} \sim \text{Target} + \text{Peak} + \text{Deaccent} + (1 | \text{Participant}) + \text{Target:Peak} + \text{Target:Deaccent} \].

ICC (indicative of the correlation of the items within a cluster) as well as marginal and conditional R-squared statistics (Nakagawa et al., 2017) were obtained using sjPlot package (Lüdecke, 2018). 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 magnitude of perceived prominence predicted by the model is shown in Figure 6. The amount of variance explained by both fixed and random effects amounted to over 60%. Fixed effects account for only about 6% of the variance. Analysis of the fixed effects suggests an association between perceived prominence and manipulations involving $R_d$ peak height and postfocal deaccentuation: increasing peak height and the steepness of deaccentuation results in an increase in syllable prominence. In other words, changing $R_d$ towards tenser phonation in the focal syllable and increasing the extent of postfocal deaccentuation by changing $R_d$ more towards laxer phonation results in an increase in the magnitude of perceived prominence of the target syllable. This effect is stronger in WAY than in YEAR. It should be noted, however, that the size of the effect is very small.

The naturalness and the confidence ratings are discussed for both tests in Section 3.4.

As mentioned earlier in the description of methodology (Section 2.2), the same 41 stimuli were presented to another group of participants ($n = 28$) in Test 2. In this test, rather than deciding on the prominence of either WAY or YEAR, participants were asked to mark the relative prominence of all the syllables in the utterance on a continuous visual analogue scale. They were also asked to rate the naturalness of the stimuli and to indicate how confident they were in their ratings.
The obtained magnitude ratings were first normalized [0, 1] to account for differences in the use of the visual analogue scale range by individual participants. These normalized values were used in the subsequent analyses.

### 3.2.1 General observations

Figure 7 (top panels) shows the mean normalized magnitude of perceived prominence of the syllables in the synthetic stimuli (blue lines = WAY-manipulated stimuli; red lines = YEAR-manipulated stimuli) relative to the baseline stimulus (black). Although the representation in Figure 7 does not allow for detailed comparison of individual types of manipulation, it is clear that practically all stimuli of the WAY-manipulated set yielded perceptual enhancement of WAY relative to the baseline. However, only some stimuli were effective in signaling prominence in the YEAR-manipulated set (so that YEAR actually sounded more prominent than WAY).

The baseline stimulus where both WAY and YEAR were prosodically flattened in terms of $f_0$, $R_d$ (see also Section 2.1.4) was perceived as having somewhat more prominent WAY. This bias towards higher prominence of WAY also affects the perception of YEAR-manipulated stimuli. This might go some way to explain why relatively few of the stimuli from the YEAR-manipulated set emerged as effective in ‘tipping the balance’ of perceived prominence from WAY to YEAR.

It is worth noting also that the prominence of the unstressed syllables before and after focally accented syllables, was adjusted by the participants to a level below that of the corresponding unstressed syllables in the baseline stimulus.

As the baseline stimulus still conveyed information on the syllable prominence (more prominent WAY than YEAR), we normalized the prominence magnitude contours further to remove the undesirable characteristics of the baseline and to be able to more clearly observe the effect of controlled $R_d$ manipulations. This was done by subtracting the prominence magnitude of the baseline from the prominence magnitude ratings of each speaker. It is clear from these normalized prominence contours (Figure 7, bottom panels) that in both way- and year- sets most of the stimuli are rated as having prominence on the target syllable (which is rated above the baseline) with the non-prominent syllable receiving prominence magnitude below the baseline.

### 3.2.2 Performance of individual stimuli

A more detailed treatment of the individual stimuli is given in Figure 8 which shows the baseline-normalized difference in perceived prominence of WAY and YEAR syllables within the same stimulus (essentially the figure captures whether and to what extent the prominence balance is tipped towards WAY or YEAR in the manipulated syllables). The bars in Figure 8 show the difference in prominence magnitude of WAY and YEAR, so when the WAY is perceived as more prominent, the values are positive and when the YEAR is more prominent the values are negative. The color coding indicates what syllables were manipulated to generate prominence.

Our initial model in the mixed-effect analysis included Manipulation and Syll as fixed predictors with by-participant random intercept $\text{diff}_\text{val} \sim \text{Syll} + \text{Manipulation} + (1|\text{Participant})$. A likelihood ratio test indicated that including Syll as a fixed factor to the Manipulation-only model significantly improved the model fit $\chi^2 (1) = 356.9$, $p < .001$ [unsurprisingly]. However, due to data sparsity this
model was found rank-deficient. We thus conducted the analysis separately for the way- and year-sets; linear mixed-effect models were fit to the baseline-normalized WAY-YEAR difference values as a dependent variable and Manipulation as the fixed factor; a by-subject intercept was included as random effect. Analyses were conducted in the R environment (R Core Team, 2019) using the lme4 \[version 1.1-29\] package (Bates et al., 2015) for model fitting. Model estimates are shown in Table 5 and Supplementary Figure 1.

As noted earlier, there is a clear difference between the effect of voice source manipulation on the relative prominence of WAY and YEAR in the two stimulus sets, with a bias towards WAY. Twelve out of twenty stimuli in the WAY-manipulated set generated prominence on WAY significantly above the baseline. Stimuli containing high peak (HP) and steep deaccentuation appear to have the largest effect on prominence (see Table 5 and Figure 1 of supplementary material).

In the YEAR-manipulated stimulus set, only six out of twenty stimuli were perceived as having YEAR more prominent than WAY. These are LP, HP, LP+Early, Steep, LP+Steep, LP+Late+Steep. Stimuli containing high peak (HP) are not predominant among those enhancing prominence for YEAR, four out of six stimuli contain low peak (LP) with or without further manipulations such as peak timing and deaccentuation. Against expectations, combined stimuli containing HP (= higher increase in phonatory tension) did not yield higher prominence enhancement compared to the LP ones.

Comparing the results of mixed-model analysis in Table 5, stimuli that yielded enhancement of perceived prominence in the target syllables in both manipulated sets (highlighted in yellow in Table 5) are: HP, LP+Early, LP+Steep, LP+Late+Steep. Apart from HP, all these stimuli involve additional adjustments of either peak timing or rate of deaccentuation. It should be noted, however, that the effect of the $R_d$ manipulations is small (low marginal $R^2$ values) accounting only for 7% (way-set) and 3% (year-set) of the variance.

Stimuli achieving significant increase in the prominence of the target syllable relative to the baseline were largely different for WAY and YEAR sets. The effect of the location of the manipulated syllable in the utterance and consequently the length of the postfocal material appears to be of importance. It is also possible that increased phonatory tension of the voice is not effective in cueing focal accentuation for syllables located towards the end of an utterance because it runs counter to the overall source deaccentuation/declination (which may be associated with laxer phonation or an increase in creaky phonation) and needs to be adjusted accordingly.

### 3.2.3 Perceived prominence and types of manipulation

Mixed model analysis was conducted to explore if and to what extent perceived prominence of the manipulated syllables can be predicted from the types of manipulation (peak magnitude, peak timing, deaccentuation). The analyses were conducted separately for way-manipulated and year-manipulated sets in the R environment (R Core Team, 2019) and followed the same procedure as described above for Test 1 (Section 3.1.2). The lmerTest package (Kuznetsova et al., 2017) 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 fit by the maximum likelihood (ML) method.
The initial model included the WAY-YEAR prominence difference as the dependent variable, Peak, Timing and Deac as fixed predictors as well as their interactions; random effects included by-subject random intercepts [diff_value~ Peak*Timing*Deac +(1|Participant)].

The final reduced model for the way-set excluded a number of non-significant interactions: [diff_value_w ~ Peak + Deac + Timing + (1 | Participant) + Peak:Deac + Deac:Timing]; for the year-set the model was not reduced. Estimated coefficients of the mixed effect models are shown in Table 6.

Fixed effects (marginal $R^2$) are very small accounting for 4% of the variance in the way-set and 2% of the variance in the year-set. Random and fixed effects (conditional $R^2$) accounted for about 59% and 51% of the variance in the way-set and year-set data respectively. Analysis of the fixed effects suggests an association between manipulations which involved peak height and postfocal deaccentuation and an increase in the perceived difference in the relative prominence of WAY and YEAR. [Peak and Deac were found to be important predictors of prominence magnitude in Test 1 also]. Two-way interactions Peak:Deac and Deac:Timing were significant in both sets; three-way interaction Peak:Deac:Timing was significant only in the year-set.

The main effects of manipulations on perceived prominence of the target syllables in the two sets are shown in Figure 9. As clear from Figure 9, the effect of manipulations varies with the location of the manipulated syllables in the utterance (and also with the length of the postfocus deaccentuation); the effect is much stronger in the WAY-manipulated utterances than in the YEAR-manipulated ones.

In WAY-manipulated utterances, as the peak magnitude increases, the perceived prominence of the syllable increases. The extent of the increase diminishes with the increase in the steepness of deaccentuation (Peak:Deac interaction). A delay in the $R_d$ peak (Timing = Late) appears to enhance the prominence (Deac:Timing interaction). In YEAR-manipulated utterances, the observed trend is different from the way-set and there is a clear three way interaction effect. As the height of the $R_d$ ‘peak’ increases, the prominence of the manipulated syllable increases, but the trend is counterbalanced and even reversed (for early peaks) with an increase in the steepness of deaccentuation (Peak:Deac:Timing interaction).

3.3 Comparing stimuli performance across the two tests

As the tasks for the two tests were rather different – the first one required a forced choice decision selecting a prominent syllable (WAY, YEAR, Neither) while the second one asked the participants to mark the prominence of all the syllables in the manipulated utterances, only very broad comparison of the performance of individual stimuli across the two tests can be made. One would expect that the stimuli most frequently selected by participants as prominent in Test 1 would also be the ones resulting in significant enhancement of prominence in Test 2. The results are very different for way- and year-sets (Supplementary Table 1). There are quite a few stimuli that were frequently selected as prominent in Test 1 and also achieved significant enhancement of prominence in Test 2 in the way-set, but none of the stimuli was as effective in the year-set. Four stimuli, HP, LP+Early, LP+Steep, LP+Late+Steep significantly enhanced the prominence of the manipulated syllable in Test 2 in both way- and year-sets, but the year-set ones were not selected in 70% of the cases in Test 1. Although this 70% cutoff is rather arbitrary, it shows again a very clear bias towards the WAY syllable.
3.4 Stimuli naturalness and confidence ratings

The overall mean naturalness ratings of the stimuli and the confidence in ratings were as follows:

Test 1 - Naturalness 43.3 (SD = 25.7, n = 597), confidence 64.7 (SD = 22.8, n = 597). Pearson’s correlation between naturalness and confidence \( r(595) = 0.22, t = 5.49, p < .001 \) (significant but weak correlation).

Test 2 – Naturalness 50.5 (SD = 21.9, n = 1148), confidence 59.1 (SD = 21.0, n = 1148). Pearson’s correlation between naturalness and confidence \( r(1145) = 0.17, t = 5.98, p < .001 \) (significant but weak correlation). Unsurprisingly, there is a significant, albeit weak, positive correlation between naturalness and confidence ratings in both tests.

Fitting a mixed-effect model to naturalness ratings in Test 1 \([\text{Naturalness} \sim \text{Target} + (1|\text{Participant})]\) showed that Target syllable is not a significant predictor of stimulus naturalness. Refitting the model to Peak*Timing*Deac as fixed predictors with a random by-participant intercept (Supplementary Figure 2) showed that naturalness ratings tend to decrease as the extent of individual manipulations increases, e.g., high peak or steep declination may signal higher prominence but may not necessarily sound natural. However, the results also showed significant Peak*Timing and Timing*Deac interaction as well as three-way Peak*Timing*Deac interaction.

4 Discussion

4.1 Overall summary of results

In this paper we describe two perception experiments exploring the perceptual salience of voice source adjustments in signaling focal prominence in the absence of \( f_0 \) variation. The sentence ‘We were away a year ago’ was analyzed using manual interactive inverse filtering and source parameterization and subsequently manipulated to flatten its prosody in terms of \( f_0 \) and \( R_d \), yielding the baseline stimulus. Stimuli were then constructed in which the voice source (the \( R_d \) parameter) was manipulated in order to achieve focal prominence on either WAY or YEAR syllables. The manipulations involved decreasing the \( R_d \) parameter (tenser voice) in the syllables WAY or YEAR and increasing it (laxer voice) in the following material, in order to mimic postfocal deaccentuation. The extent of \( R_d \) manipulation varied in terms of the height of \( R_d \) ‘peak’, its location within the syllable (early or late) and extent of deaccentuation. \( f_0 \) remained constant across the stimulus set. The stimuli were presented to participants in two listening tests. In Test 1, the participants were asked to identify which of WAY or YEAR was the most prominent, with a Neither option where listeners could not decide between them. In Test 2, the participants were asked to indicate the relative prominence of all the syllables in the utterance on a visual analogue scale.

Our main research questions were: (1) To what extent can such source manipulations induce the perception of focal accent on one or other syllable? (2) Which of the source manipulations (or which combinations of source manipulations) are most effective in cueing focal accentuation? Our results suggest that voice source (here \( R_d \) parameter) manipulation, even in the absence of \( f_0 \) salience, produced the desired effect of shifting focal prominence to the manipulated syllable, although in this study there was a clear bias towards the WAY syllable. Not all manipulations or their combinations achieved cueing of prominence and the stimuli effective in cueing prominence were not identical in the two tests. Increasing the height of the \( R_d \) ‘peak’ (changing \( R_d \) towards tenser phonation in the focal syllable) and increasing the extent of postfocal deaccentuation by changing \( R_d \) towards laxer phonation appear to be particularly effective and resulted in an increase in the magnitude of...
perceived prominence of the target syllable. This effect is stronger in WAY than in YEAR. The naturalness of the stimuli appears to decrease with an increase in the extent of source manipulation.

4.2 Expectations/hypotheses against findings

Our expectation was that the syllables WAY and YEAR in the sentences where the voice source for those syllables and for the following material was systematically manipulated would tend to be identified as more prominent. We also expected that the degree of prominence perceived on the targeted syllable would correlate with the extent of the source manipulation carried out, e.g., the higher the $R_d$ ‘peak’ (and the tenser the voice), the more prominent the syllable would be rated or conversely, greater $R_d$ deaccentuation would be correlated with greater focal prominence on the preceding syllable. Furthermore, it was expected that the combined manipulations would produce higher prominence ratings compared to individual manipulations. We expected broad similarities across the two tests. Our initial hypotheses did not include any predictions about the effect of the location of the potentially accentable syllable in the utterance.

4.2.1 Expectation: Syllables where systematic manipulation was done are perceived as prominent

As expected, syllables where systematic manipulations were done were perceived as prominent by the participants of the listening tests, but there was a strong bias towards WAY: all stimuli in the way-set were perceived with prominence on the target syllable in both tests and above the baseline. This was not the case for the year-set: in Test 1, the perception of YEAR prominence was closer to chance, and none of the stimuli were perceived as prominent by 70% of participants or more. In test 2, only 6/20 stimuli were perceived as having higher prominence on YEAR than WAY relative to the baseline (Supplementary Table 1).

4.2.2 Expectation: The degree of prominence perceived on the targeted syllable correlates with the extent of manipulation; combined manipulations are more effective

The results (Test 2, Figure 9) suggest that the extent of manipulation (peak height, steepness of deaccentuation) has an effect on the magnitude of perceived prominence, but this effect is different for the two target syllables. In WAY-manipulated utterances, perceived prominence increased with the increase in the $R_d$ ‘peak’ and was further enhanced by a delay in the $R_d$ ‘peak’. Addition of a steeper deaccentuation, however, diminished this effect. In YEAR-manipulated utterances, the observed trend is different from the way-set and there is a clear three-way interaction effect of Peak, Timing and Deaccentuation. As the height of the $R_d$ ‘peak’ increases, the prominence of the manipulated syllable increases, but the trend is counterbalanced and even reversed (for early peaks) with an increase in the steepness of deaccentuation (Peak:Deac:Timing interaction). It is worth noting here that the extent of manipulation appears to have a negative effect on the naturalness of the stimuli.

Although the results seem to suggest that combining more extreme Peak and Deaccentuation enhances prominence cuing, there is no strong consistent evidence in our data that combined manipulations are more effective than individual ones. For example, HP, HP+Early and HP+Late+Shallow were equally effective in signaling prominence of WAY in Test 2 (the difference between them is not significant). In the year-set, the Steep stimulus was as effective as LP+Late+Steep or LP+Early. In fact, HP was the only stimulus that was consistently effective in signaling prominence across the two sets. The timing of the $R_d$ ‘peak’ appear to be of perceptual consequence in our experiments only as an interaction effect.
4.2.3 Cuing of focal accentuation can vary depending on the location of the focal syllable in the utterance

In our experiments the two potentially accentable syllables were manipulated in order to evoke prominence. We made no specific a priori predictions about the impact of the location of the manipulated syllable within an utterance on its perceived prominence.

Stimuli achieving significant increase in the prominence of the target syllable relative to the baseline were largely different for WAY and YEAR (Table 5). It is clear in these data that the cueing of focal accentuation can vary depending on its location in the utterance. In the non-final position (i.e. WAY), even relatively small changes in the source parameter values (e.g., LP+Early, LP+Steep) appear to make a difference, and can tip the balance in terms of where focal accent is likely to be perceived. It is also clear that there is a synergy between the local prominence on the syllable (HP) and deaccentuation in the postfocal material.

In the final accentable syllable (YEAR) the findings were not symmetrical. Here Steep and LP ‘simple’ manipulations significantly enhance the prominence ratings, something not observed in the way-set. While HP on its own is effective in this set as well, combinations including low peak and steep declination are more effective than combinations with high peak.

As mentioned earlier, the effect of the location of the manipulated syllable in the utterance and consequently the length of the postfocal material appears to be of importance. It is also possible that increased phonatory tension of the voice (HP) is not effective in cueing focal accentuation for syllables located towards the end of an utterance (the year-set) because it runs counter to the overall source deaccentuation/declination trend associated with increasingly laxer phonation. In future work this overall source declination trend needs to be included in synthetic stimuli. There is similarity here to \( f_0 \) peaks and the ‘compensation for declination’ (Gussenhoven & Rietveld, 1998; Pierrehumbert, 1979; Terken & Hermes, 2000): for the \( f_0 \) peak to be perceived as having the same pitch as the previous one it needs to be lower.

It is likely that the differences observed here between the final (YEAR) and non-final (WAY) syllables have to do with what was not included in these tests, i.e. manipulations to \( f_0 \). The \( f_0 \) was kept constant in these stimuli as the objective was to ascertain the role of other voice source features. However, \( f_0 \) movement co-occurs with the kinds of source features implemented here and it is very likely that \( f_0 \) movement is far more crucial in final than in non-final syllables. In a production study of focus (Yanushevskaya et al., 2010) an \( f_0 \) fall was found in both WAY and YEAR syllables when focally accented, but the fall was greater and more rapid in YEAR. A further study of source correlates of accentuation (Ní Chasaide et al., 2013) indicated that while accented syllables in non-final position may, but need not, exhibit \( f_0 \) movement, a sharp \( f_0 \) fall always characterized the final accented syllable. To the extent that this fall is missing in the present stimuli, it is likely to militate strongly against the perception of greater prominence on YEAR, regardless of the source changes that occur.

Apart from the position in the utterance, other factors may be relevant. Vowels in the target syllables in the stimuli are of different quality, which may have an impact on perceived prominence. Open vowels have inherently greater intensity and longer duration than close vowels (van Heuven, 2014). This could further explain the bias towards WAY we observed in our results.
4.2.4 Findings across the tests are not identical: test setup has an impact

The stimuli were presented to participants in two different online tests, and the tasks were different: in Test 1, the participants had to select the prominent syllable (forced choice) and estimate its prominence, in Test 2 they were required to mark the prominence of all the syllables in an utterance on a visual analogue scale. There are certain similarities in the findings: the same types of manipulation were effective in both tests (though not in both sets); in both tests we see a strong bias towards WAY; the naturalness and confidence ratings of the stimuli were largely similar (Supplementary Table 1). However, against expectations, the stimuli most frequently selected by participants as prominent in Test 1 were not necessarily the ones yielding notable enhancement of prominence in Test 2. As mentioned earlier, stimuli LP+Early, LP+Steep, LP+Late+Steep significantly enhanced the prominence of manipulated syllable in Test 2 in both way- and year- sets, however, these stimuli in the year-set were not selected by 70% of the participants in Test 1.

These differences in the findings suggest that the context of the test has an effect on the listener prominence judgement.

4.3 Strengths – what this paper contributes

This paper is the first to look at the perceptual salience of source variation when f0 is kept constant. Our earlier production and perception studies (Gobl et al., 2002; Gobl & Ni Chasaide, 2003b; Ryan et al., 2003; Yanushevskaya et al., 2018) involved the analysis and synthesis of many voice source parameters, many of which covary in natural speech production. Controlling such an array of complex parameters in synthesis is difficult and require expert knowledge, and it would be therefore desirable to achieve control of voice quality modulation using a smaller set of parameters. This work is a contribution to our understanding of how the Rd parameter might be used as a control parameter in synthesis. This research contributed to the development of a system, described in (Murphy et al., 2021) that allows the user to manually manipulate the voice quality dimension of prosody of a synthetic utterance using a graphical interface for use in voice and prosody research as well as in various applications involving synthetics speech (educational games, assistive technology). The paper further contributes to the development of a model of voice source modulation in linguistic prosody (focus, prominence and deaccentuation).

4.4 Limitations

To keep the number of manipulations manageable, synthetic stimuli were created adopting a rather simplistic modelling approach. The parameter contours were stylized using linear interpolation, the overall source declination was not included and the heights of the later peak was not adjusted accordingly; duration of the syllables were not adjusted to account for pre-pausal lengthening and the vowel quality in the target syllables was different. For example, WAY is slightly longer than YEAR [8 ms, below JND threshold of 10 ms (Plack, 2018)] but is located earlier in the utterance (YEAR needs to be longer rather than shorter to account for the overall lengthening of the syllables towards the end of the utterance). Furthermore, the [eɪ] vowel in WAY is more open and its intensity is intrinsically higher than that of a closer vowel in YEAR. These differences might have contributed to the strong bias towards prominence on WAY that we observed in our data for both way- and year-sets.

There were sources of variation in our data that were not accounted for by the models we used. The length of postfocal material and the extent of deaccentuation were different for the syllables located earlier and later in the utterance. These were not possible to control for in the current experimental
setup, but their influence on perceived prominence of manipulated syllables need to be explored separately in a follow-up study.

The aim of the experiments was to establish perceptual salience of source modulations in the absence of $f_0$ variation. In speech, $f_0$ and other source parameters interact (Fant, 1997; Fant & Kruckenberg, 2007; Ní Chasaide et al., 2011); this was ignored in the current experiment.

Perception experiments were conducted with a relatively small number of participants. The certainty of the point estimates of the magnitude data under the models in our study is rather low as shown by the wide 95% CI, so only tentative conclusions can be made. A follow-up confirmatory study is necessary.

The naturalness ratings of the stimuli were not very high and were negatively correlated with the extent of manipulations. It may be the case that the target syllables were marked as prominent because they sounded less natural. As mentioned earlier, the low naturalness ratings are likely to be related to the extent of deviation from the original parameter values in combination with the absence of $f_0$ manipulations. Source modulation in real speech, although separately controllable, typically co-occur with changes in $f_0$. The discussion of naturalness is relevant: the higher extent of manipulation appears to have a negative effect on naturalness. This suggests that extreme manipulations are less desirable if it is possible to achieve prominence with relatively moderate manipulations.

5 Conclusions

Our exploratory study indicates that increasing phonatory tension in the focal syllable and reducing it in post-focal material by manipulating the global waveshape parameter $R_d$ can be effective in cuing focal prominence in synthetic speech in the absence of $f_0$ modulation. It further suggests that having a source prominence peak (tenser phonation) on the focally accented syllable may work synergistically with a degree of source deaccentuation in the postfocal material.

The manipulations that induced the perception of focal accentuation in the non-final syllable (WAY) had much less effect on the syllable located later in the utterance (YEAR), where focal accentuation was not well cued. The cueing of focal prominence may depend on its location in the utterance, and that in the case of the final accented syllable $f_0$ movement (not included) is a necessary component.

However, conclusions based on a single study can only be tentative (Nicenboim et al., 2018) and confirmatory follow-up studies need to be conducted. As a next step in these studies, we hope to look at the interplay of source parameters with $f_0$ in final and non-final syllables, and also the effects of deaccentuation when the postfocal tail is longer. Future work will also control for vowel quality and duration in focally accented syllables as well as for the overall source declination in the course of an utterance. The effect of the manipulations of $R_d$ parameter in our experiments is rather low (2-6% of the variance, as shown by marginal $R^2$ values). In natural speech, $f_0$ and other source parameters work synergistically. Future work is required to test to what extent source parameters such as $R_d$ enhance prominence, when manipulated together with $f_0$.

Our subsequent work reported in (Murphy et al., 2018) made use of listener-driven decisions. The same approach needs to be applied here to establish perception-driven optimal prominence settings. Adjusting local manipulations of the target syllables to control for the overall source declination as well as pre-focal and post-focal material is likely to improve the naturalness of the stimuli. This will also bring us closer to formulating voice prominence model.
Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

All authors contributed to conception and design of the study. AM generated synthetic stimuli. IY and AM collected the listening test data. IY performed inverse filtering and source parameterisation as well as the statistical analyses and wrote the first draft of the manuscript. AM, CG and ANC wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This research was carried out in the Róbóglór-ABAIR project, which is supported by An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán, with funding from the National Lottery, as part of the Straitéís 20 Blaíon don Ghaeilge.

Acknowledgments

We thank Christopher La Pat for assistance with the setup of the listening tests. We thank the Editor and Reviewers for their helpful and detailed comments. An earlier version of this research was presented at Interspeech 2016 (Yanushevskaya, Murphy, et al., 2016).

Data Availability Statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References


$R_d$ parameter in signaling focal prominence


**Ra parameter in signaling focal prominence**


$R_d$ parameter in signaling focal prominence


**Rd parameter in signaling focal prominence**


  
  https://doi.org/10.1098/rsif.2017.0213


  
  https://doi.org/https://doi.org/10.1016/j.wocn.2018.06.001


  
  https://doi.org/https://doi.org/10.4324/9781315208145


**Rad parameter in signaling focal prominence**

http://link.aip.org/link/?JAS/100/2471/1


http://link.aip.org/link/?JAS/89/1768/1

1093 http://dx.doi.org/10.1121/1.401019


http://link.aip.org/link/?JAS/99/3782/1


https://doi.org/http://dx.doi.org/10.1016/j.wocn.2005.06.004


Rd parameter in signaling focal prominence


Figure captions

Figure 1. The LF-model extended to include the $R_d$ parameter (adapted from Fant 1995).

Figure 2. Schematic of types of manipulation in the synthesized stimuli: (a) peak magnitude; (b) peak timing (rate of change); (c) post-focal deaccentuation.

Figure 3. Experimental interface for Test 1 (left panel) and Test 2 (right panel).

Figure 4. Frequencies with which syllables WAY, YEAR or neither were identified as prominent in the synthetic stimuli. LP = low peak, HP = high peak, Late = late peak, Early = early peak, Steep = steep deaccentuation, Shallow = shallow deaccentuation (see also Table 1). Red dotted lines indicate results for the Baseline stimulus.

Figure 5. Perceived prominence of manipulated syllables in Test 1. Shown are prominence magnitude mean and 95% CI of stimuli with different types of manipulation in Test 1 (panel A) and general spread of point estimates (means) (panel B).

Figure 6. Predicted values of perceived prominence (Test 1). Shaded areas represent 95% confidence intervals.

Figure 7. Perceived prominence of all syllables in synthetic stimuli in Test 2. Top panels: the contours for individual stimuli averaged across participants; baseline stimulus is shown in black. Bottom panels: same contours normalized to the baseline.

Figure 8. The difference in perceived prominence of WAY and YEAR syllables within the same stimulus, baseline-normalised (Test 2).

Figure 9. Predicted values of difference in the relative prominence of WAY and YEAR in the same utterance (Test 2). Shown are the interactions of Peak, Timing, Deaccentuation types of manipulations. Shaded areas represent 95% confidence intervals.
Table 1. Types of manipulation and their combinations in synthetic stimuli.

<table>
<thead>
<tr>
<th>Stimulus code</th>
<th>Peak magnitude</th>
<th>Peak timing</th>
<th>Deaccentuation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Baseline</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Peak magnitude</td>
<td>LP low</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>HP high</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Peak magnitude + peak timing</td>
<td>LP+Early low</td>
<td>early</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LP+Late low</td>
<td>late</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>HP+Early high</td>
<td>early</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>HP+Late high</td>
<td>late</td>
<td>0</td>
</tr>
<tr>
<td>Deaccentuation</td>
<td>Shallow 0</td>
<td>0</td>
<td>shallow</td>
</tr>
<tr>
<td></td>
<td>Steep 0</td>
<td>0</td>
<td>steep</td>
</tr>
<tr>
<td>Peak magnitude + deaccentuation</td>
<td>LP+Shallow low</td>
<td>0</td>
<td>shallow</td>
</tr>
<tr>
<td></td>
<td>LP+Steep low</td>
<td>0</td>
<td>steep</td>
</tr>
<tr>
<td></td>
<td>HP+Shallow high</td>
<td>0</td>
<td>shallow</td>
</tr>
<tr>
<td></td>
<td>HP+Steep high</td>
<td>0</td>
<td>steep</td>
</tr>
<tr>
<td>Peak magnitude + peak timing + deaccentuation</td>
<td>LP+Early+Shallow low</td>
<td>early</td>
<td>shallow</td>
</tr>
<tr>
<td></td>
<td>LP+Late+Shallow low</td>
<td>late</td>
<td>shallow</td>
</tr>
<tr>
<td></td>
<td>HP+Early+Shallow high</td>
<td>early</td>
<td>shallow</td>
</tr>
<tr>
<td></td>
<td>HP+Late+Shallow high</td>
<td>late</td>
<td>shallow</td>
</tr>
<tr>
<td></td>
<td>LP+Early+Steep low</td>
<td>early</td>
<td>steep</td>
</tr>
<tr>
<td></td>
<td>LP+Late+Steep low</td>
<td>late</td>
<td>steep</td>
</tr>
<tr>
<td></td>
<td>HP+Early+Steep high</td>
<td>early</td>
<td>steep</td>
</tr>
<tr>
<td></td>
<td>HP+Late+Steep high</td>
<td>late</td>
<td>steep</td>
</tr>
</tbody>
</table>

Table 2. Summary confusion matrix of perception of the stimuli in Test 1 (count, %).

<table>
<thead>
<tr>
<th>Selected as prominent</th>
<th>WAY-manipulated</th>
<th>YEAR-manipulated</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAY</td>
<td>369 64%</td>
<td>216 37%</td>
<td>11 38%</td>
</tr>
<tr>
<td>YEAR</td>
<td>100 17%</td>
<td>228 39%</td>
<td>11 38%</td>
</tr>
<tr>
<td>Neither</td>
<td>111 19%</td>
<td>136 23%</td>
<td>7 24%</td>
</tr>
<tr>
<td>Total</td>
<td>580 100%</td>
<td>580 100%</td>
<td>29 100%</td>
</tr>
</tbody>
</table>
Table 3. Results of multinomial logistic regression analysis. Significant predictors shown in bold type.

<table>
<thead>
<tr>
<th>(a) WAY-manipulated stimuli</th>
<th>B</th>
<th>SE</th>
<th>Sig.</th>
<th>Odds ratio</th>
<th>95% CI for odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Way vs. Neither</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.30</td>
<td>0.31</td>
<td>.330</td>
<td>0.73</td>
<td>0.40 - 1.35</td>
</tr>
<tr>
<td>Peak</td>
<td>0.66</td>
<td>0.18</td>
<td>.000</td>
<td>1.93</td>
<td>1.36 - 2.74</td>
</tr>
<tr>
<td>Timing</td>
<td>0.01</td>
<td>0.14</td>
<td>.920</td>
<td>1.01</td>
<td>0.77 - 1.33</td>
</tr>
<tr>
<td>Deac</td>
<td>0.61</td>
<td>0.14</td>
<td>.000</td>
<td>1.84</td>
<td>1.38 - 2.43</td>
</tr>
<tr>
<td>Year vs. Neither</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.20</td>
<td>0.36</td>
<td>.576</td>
<td>1.22</td>
<td>0.60 - 2.51</td>
</tr>
<tr>
<td>Peak</td>
<td>-0.18</td>
<td>0.22</td>
<td>.413</td>
<td>0.83</td>
<td>0.53 - 1.29</td>
</tr>
<tr>
<td>Timing</td>
<td>0.03</td>
<td>0.17</td>
<td>.847</td>
<td>1.03</td>
<td>0.73 - 1.33</td>
</tr>
<tr>
<td>Deac</td>
<td>-0.14</td>
<td>0.18</td>
<td>.433</td>
<td>0.86</td>
<td>0.60 - 1.24</td>
</tr>
</tbody>
</table>

(b) YEAR-manipulated stimuli

| Way vs. Neither             |       |       |      |            |                      |
| Intercept                   | 0.16  | 0.30  | .594 | 1.17       | 0.65 - 2.11          |
| Peak                        | 0.05  | 0.17  | .772 | 1.05       | 0.74 - 1.47          |
| Timing                      | -0.01 | 0.13  | .954 | 0.99       | 0.75 - 1.30          |
| Deac                        | 0.22  | 0.14  | .104 | 1.25       | 0.95 - 1.64          |
| Year vs. Neither            |       |       |      |            |                      |
| Intercept                   | -0.29 | 0.31  | .355 | 0.74       | 0.40 - 1.38          |
| Peak                        | 0.61  | 0.17  | .000 | 1.85       | 1.30 - 2.62          |
| Timing                      | -0.002| 0.13  | .985 | 0.99       | 0.76 - 1.30          |
| Deac                        | -0.03 | 0.13  | .771 | 0.96       | 0.73 - 1.25          |

Table 4. Estimated coefficients, confidence intervals and t values for the mixed effect model fitted to the perceived prominence magnitude data (Test 1).

<table>
<thead>
<tr>
<th>Predictors</th>
<th>β₀</th>
<th>CI</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>41.35</td>
<td>31.25 – 51.45</td>
<td>8.03</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Target [year]</td>
<td>-0.55</td>
<td>-5.06 – 3.96</td>
<td>-0.24</td>
<td>.811</td>
</tr>
<tr>
<td>Peak</td>
<td>3.60</td>
<td>-0.62 – 7.81</td>
<td>1.67</td>
<td>.094</td>
</tr>
<tr>
<td>Deac</td>
<td>0.28</td>
<td>-4.45 – 5.00</td>
<td>0.11</td>
<td>.909</td>
</tr>
<tr>
<td>Target [year] * Deac</td>
<td>-3.90</td>
<td>-7.23 – -0.57</td>
<td>-2.29</td>
<td>.022</td>
</tr>
<tr>
<td>Peak * Deac</td>
<td>3.39</td>
<td>0.61 – 6.16</td>
<td>2.39</td>
<td>.017</td>
</tr>
</tbody>
</table>

Random effects

<table>
<thead>
<tr>
<th></th>
<th>β₀</th>
<th>n = 29</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC Participant</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Marginal R² / Conditional R²</td>
<td>.060/ .607</td>
<td></td>
</tr>
</tbody>
</table>
Table 5. Estimated coefficients, confidence intervals and t values for the mixed effect model with Manipulation as a fixed factor fitted to the perceived prominence magnitude data (baseline-normalised) in Test 2. Manipulations that yielded significant shifts in prominence relative to the baseline are shown in bold type and are highlighted in yellow if effective for both sets.

<table>
<thead>
<tr>
<th>WAY</th>
<th>YEAR</th>
<th>Predictors</th>
<th>β₀</th>
<th>CI</th>
<th>t</th>
<th>p</th>
<th>β₀</th>
<th>CI</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (=Baseline)</td>
<td>0</td>
<td>-15.71 – 15.71</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>-15.89 – 15.89</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LP</td>
<td>14.07</td>
<td>6.48 – 21.65</td>
<td>2.26</td>
<td>.025</td>
<td>-1.98</td>
<td>.050</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP</td>
<td>31.13</td>
<td>16.86 – 45.40</td>
<td>4.29</td>
<td>&lt;.001</td>
<td>-1.99</td>
<td>.049</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>LP+Late</td>
<td>10.30</td>
<td>-3.97 – 24.56</td>
<td>2.42</td>
<td>.157</td>
<td>2.40</td>
<td>.016</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Late</td>
<td>19.72</td>
<td>5.00 – 33.54</td>
<td>2.65</td>
<td>.008</td>
<td>-7.08</td>
<td>.004</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Early</td>
<td>34.71</td>
<td>20.44 – 48.98</td>
<td>4.78</td>
<td>&lt;.001</td>
<td>-12.51</td>
<td>.001</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>Shallow</td>
<td>-2.62</td>
<td>-16.89 – 11.65</td>
<td>-0.36</td>
<td>.719</td>
<td>-6.18</td>
<td>.076</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>Steep</td>
<td>4.51</td>
<td>-9.76 – 18.78</td>
<td>0.62</td>
<td>.535</td>
<td>-21.70</td>
<td>.007</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>LP+Shallow</td>
<td>2.23</td>
<td>-12.04 – 16.50</td>
<td>0.31</td>
<td>.759</td>
<td>-6.72</td>
<td>.042</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>LP+Steep</td>
<td>22.40</td>
<td>8.13 – 36.67</td>
<td>3.08</td>
<td>&lt;.001</td>
<td>-15.94</td>
<td>.047</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Shallow</td>
<td>11.23</td>
<td>-3.04 – 25.50</td>
<td>1.55</td>
<td>.123</td>
<td>-6.65</td>
<td>.040</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Shallow</td>
<td>6.90</td>
<td>2.80 – 11.00</td>
<td>1.20</td>
<td>.246</td>
<td>-6.14</td>
<td>.044</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Early</td>
<td>18.70</td>
<td>4.43 – 32.97</td>
<td>2.57</td>
<td>.010</td>
<td>-2.20</td>
<td>.027</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>LP+Shallow</td>
<td>17.59</td>
<td>3.33 – 31.86</td>
<td>2.42</td>
<td>.016</td>
<td>-6.40</td>
<td>.425</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>LP+Late</td>
<td>27.78</td>
<td>13.51 – 42.05</td>
<td>3.82</td>
<td>&lt;.001</td>
<td>-22.02</td>
<td>.006</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Late</td>
<td>18.41</td>
<td>4.14 – 32.68</td>
<td>2.55</td>
<td>.012</td>
<td>-15.05</td>
<td>.001</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Late</td>
<td>14.50</td>
<td>0.23 – 28.77</td>
<td>1.20</td>
<td>.046</td>
<td>-6.14</td>
<td>.774</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Late</td>
<td>35.71</td>
<td>21.44 – 49.98</td>
<td>4.92</td>
<td>&lt;.001</td>
<td>-5.09</td>
<td>.525</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
<tr>
<td>HP+Late</td>
<td>26.42</td>
<td>12.15 – 40.69</td>
<td>3.64</td>
<td>&lt;.001</td>
<td>-13.29</td>
<td>.098</td>
<td>2.21 – 7.45</td>
<td>0.031</td>
<td>2.40</td>
<td>.016</td>
</tr>
</tbody>
</table>

Random effects

| ICC Participant | 0.59 | 0.51 |
| N participant   | 28   | 28   |
| Observations    | 588  | 588  |
| Marginal R²     | .070 | .031 |
| Conditional R²  | .616 | .525 |

Table 6. Estimated coefficients, confidence intervals and t values for the mixed effect model with types of manipulation as fixed factors fitted to the magnitude of to the perceived prominence magnitude data (baseline-normalised) in Test 2.

<table>
<thead>
<tr>
<th>WAY</th>
<th>YEAR</th>
<th>Predictors</th>
<th>β₀</th>
<th>CI</th>
<th>t</th>
<th>p</th>
<th>β₀</th>
<th>CI</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-6.20</td>
<td>-20.41 – 8.01</td>
<td>-0.86</td>
<td>.392</td>
<td>2.34</td>
<td>0.33</td>
<td>-16.11 – 16.29</td>
<td>0.33</td>
<td>0.742</td>
<td></td>
</tr>
<tr>
<td>Peak</td>
<td>16.49</td>
<td>11.32 – 21.65</td>
<td>6.27</td>
<td>.001</td>
<td>5.62</td>
<td>1.99</td>
<td>0.07 – 11.18</td>
<td>1.99</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>Deac</td>
<td>8.66</td>
<td>2.80 – 14.51</td>
<td>2.90</td>
<td>.004</td>
<td>7.48</td>
<td>2.33</td>
<td>1.18 – 13.78</td>
<td>2.33</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>-3.84</td>
<td>-8.62 – 0.94</td>
<td>-1.58</td>
<td>.115</td>
<td>-17.64</td>
<td>2.13</td>
<td>-33.91 – 1.38</td>
<td>-2.13</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>Peak*Deac</td>
<td>-6.45</td>
<td>-10.45 – -2.45</td>
<td>-3.17</td>
<td>.002</td>
<td>-5.31</td>
<td>-2.42</td>
<td>-9.61 – 1.00</td>
<td>-2.42</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>Peak*Timing</td>
<td>6.56</td>
<td>-3.72 – 16.85</td>
<td>1.20</td>
<td>.211</td>
<td>6.65</td>
<td>3.00</td>
<td>6.65 – 31.85</td>
<td>3.00</td>
<td>0.003</td>
<td></td>
</tr>
</tbody>
</table>

Random effects

| ICC Participant | 0.57 | 0.50 |
| N participant   | 28   | 28   |
| Observations    | 588  | 588  |
| Marginal R²     | .041 | .019 |
| Conditional R²  | .586 | .513 |
$R_d$ parameter in signaling focal prominence

\begin{align*}
U_i(t) & \quad U_p \\
T_a = U_p/E_e & \quad t
\end{align*}

**Voice source parameters**

$R_d = (T_d/T_0) \ (1/0.11)$

$= (U_p/E_e) \ (f_0/0.11)$

$R_a = T_a/T_0$

$R_g = T_0/(2T_p)$

$R_k = (T_e-T_p)/T_p$

$f_0 = 1/T_0$

---

**Figure 1**

---

**Figure 2**
\( R_d \) parameter in signaling focal prominence

Figure 3

Figure 4
$R_d$ parameter in signaling focal prominence

**Figure 5**

**Figure 6**
$R_d$ parameter in signaling focal prominence

Figure 7
$R_d$ parameter in signaling focal prominence

Figure 8
$R_d$ parameter in signaling focal prominence

**Figure 9**
Supplementary Figure 10. Mixed effect model predictions of WAY-YEAR difference with Manipulation as a fixed factor and by-subject random intercept. 0=Baseline. As the model was fit to WAY-YEAR difference, negative values (in red) correspond to the cases when YEAR is perceptually more prominent than WAY.

Supplementary Figure 11. Model-predicted stimuli naturalness rating in Test 1. Main interactions (panel A) and model coefficients (panel B).
**Supplementary Table 7.** Marked with “x” are the stimuli deemed prominent by 70% or more of participants in Test 1 and the stimuli which significantly enhanced perceived prominence in Test 2 relative to Baseline. In bold shown are the cases where the manipulations were effective in both way- and year-sets in the same test.

<table>
<thead>
<tr>
<th></th>
<th>Way-manipulated</th>
<th>Year-manipulated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test 1</td>
<td>Test 2</td>
</tr>
<tr>
<td>LP</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>HP</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>LP+Late</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LP+Early</strong></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>HP+Late</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>HP+Early</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Shallow</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steep</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>LP+Shallow</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td><strong>LP+Steep</strong></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>HP+Shallow</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>HP+Steep</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>LP+Early+Shallow</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LP+Early+Steep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LP+Late+Shallow</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LP+Late+Steep</strong></td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

1264

1265