INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Perceptual salience of voice source parameters in signaling focal prominence

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

Phonetics and Speech Laboratory, Trinity College Dublin, Ireland

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

Abstract

This paper describes listening tests investigating the perceptual role of voice source parameters (other than F0) in signaling focal prominence. Synthesized stimuli were constructed on the basis of an inverse filtered utterance 'We were away a year ago'. Voice source parameters were manipulated in the two potentially accentable syllables WAY and YEAR (in terms of the absolute magnitude and alignment of peaks) and to provide source deaccentuation of post-focal material. Participants in the first listening test were asked to decide whether the syllable WAY, YEAR or neither was deemed the most prominent: judgments on the degree of prominence and naturalness were also indicated on a continuous visual analogue scale. In the second test listeners indicated the degree of prominence for every syllable in the phrase. For WAY, voice source manipulations can cue focal accentuation, and both the magnitude of the source manipulation of the syllable and the presence of source deaccentuation contribute to the effect. However, for YEAR, listeners' perception of focal accentuation tended to show relatively minor increases in perceived prominence regardless of the source manipulations involved. It therefore appears that the source expression of focus is sensitive to the location of focus in the intonational phrase.

Index Terms: voice source, RD, focus, prominence, accentuation, perception

1. Introduction

Past production studies have looked at the role of the voice source as part of sentence prosody, and have shown that voice source parameters are involved in the realization of accentuation [1], focus [2] and declination [3]. The picture emerging is that prosody entails the modulation of the entire voice source (including F0) and that the different parameters appear to work synergistically in contributing to the realization of prominence, deaccentuation, etc. In the study on accentuation [1] it was clear that even in the absence of F0 salience, other voice source parameters appear to be responsible for the signaling of prominence. It can be noted that, although F0 and source parameters often covary, they can also be controlled independently of each other.

In this paper, we set out to elucidate the perceptual importance that may attach to the kinds of voice source adjustments which we have observed in sentences with variable location of focal accent. We further explore whether such voice source adjustments on their own might be capable of shifting the perception of the location of focal accent within the sentence.

There have been many experimental studies demonstrating the role of F0 peaks [4], [5], [6], [7], [8], [9] in the realization of prominence, accentuation and focus but there is little on the perceptual role of voice source adjustments other than F0.

In this study a recording of the 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 following part of the utterance. The questions we set out to answer were: (1) Can such source manipulation induce the perception of focal accent on one or other syllable? (2)Which of the source manipulations (or which combinations of source manipulations) were most effective in cueing focal accentuation?

In these experiments, F0 did not vary across the stimulus set. This is not to suggest that F0 does not play a major role in cueing focus, but rather represents an attempt to explore how voice factors other than F0 might be contributing, and to see whether source variations alone (without F0 variation) can alter the perception of where the focal accent lies in a phrase. Note that the extent of source variation used in these stimuli falls well within the ranges observed in production studies.

2. Material: synthetic stimuli

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, and was recorded as part of another study, where further versions of the sentence with a focal accent on the syllables WAY and YEAR were also obtained and source characteristics analyzed [10]. The utterance was manually inverse filtered using interactive inverse filtering software [11], [12]. Voice source parameterization was subsequently conducted using the Liljencrants-Fant (LF) model [13]. In the construction of the synthesized stimuli only one parameter was directly manipulated: the global waveshape parameter RD [14], [15].

The RD parameter is derived from F0, EE and UP as follows: $(1/0.11)\times(F0\cdot UP/EE)$, where EE is the excitation strength (measured as the negative amplitude of the differentiated glottal flow at the time point of maximum waveform discontinuity) and UP is the peak flow of the glottal pulse. Note that UP/EE is equivalent to the glottal pulse declination time during the closing phase of the glottal cycle. The scale factor (0.11^{-1}) makes the numerical value of RD equal to the declination time in milliseconds when F0 is 110 Hz [14].

Variation in RD tends to reflect voice source variation along the tense-lax continuum; the values typically range between 0.5 (tense voice) to 2.5 (breathy voice). As our earlier

http://dx.doi.org/10.21437/Interspeech.2016-1160

analyses of the speaker used here suggest shifts towards tenser phonation in focally accented syllables [2], [16] and towards laxer phonation in the post-focal material, the adjustments made to mimic these effects in our stimuli involved lowering the values of RD in the potentially accented syllables and raising it in the post-focal part of the utterance. Note that increased phonatory tension tends to correspond to a drop in RD, but that for the purpose of this paper and the illustration in Figure 1, we refer to such RD drops as peaks, as it seems intuitively easier for a reader to associate increased phonatory tension with positive values.

In the synthesis manipulations, F0 and UP were kept constant, which means that changes in RD were reflected by changes in EE. By changing RD, other parameters of the glottal source such as RA and RK also vary, and these changes can be predicted from RD. (For a description of the various glottal parameters, see [17]). To synthesize the LF model glottal waveform, data for the full set of LF model parameters are required and were obtained from RD using the parameter correlations presented in [14] (see also [18]).

In the 'baseline' stimulus, the values of F0, RD and EE were first set to the global average values across the utterance (F0 = 120 Hz, RD = 0.86, EE = 69.8 dB). As the overall impression of this stimulus was that it sounded rather tense, the values of F0 and RD were adjusted to make it more lax and improve the naturalness: F0 was increased by 5 percent to 127 Hz and RD was increased by 50 percent to 1.3. These changes also resulted in a lowering of EE to 67.2 dB.

This version of the utterance served as the baseline for further manipulations involving the magnitude and alignment of peaks (located at the midpoint of the vowels in the syllables WAY and YEAR) as well as deaccentuation in the post-focal material. These manipulations are described in the following sections (see also schematic in Figure 1). The ranges of values used in the manipulations were based on the voice analysis of the speaker in earlier production studies mentioned above [2] [19] [16]. Note that F0 values were not manipulated and were kept constant in all the syllables of the stimuli.

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

Peak alignment: Versions of the stimuli were also generated, where peak alignment was changed relative to the vowel midpoints in the syllables WAY and YEAR. Two peak alignment settings were used, early peak and late peak; the values were shifted by 20% relative to the duration of the vowel. Early peak corresponds to faster increase to 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 1). These manipulations were added, as earlier studies of focal accentuation [20], [2], [19] 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, shallow deaccentuation and steep deaccentuation. Note that for the WAY-manipulated sentences, deaccentuation pertains to the entire sequence 'a year ago', whereas for the sentence where YEAR is manipulated, deaccentuation is necessarily limited to the syllables of 'ago'.

The RD values were as follows: no deaccentuation, final RD value = 1.3; shallow deaccentuation, final RD value = 1.1 (deaccentuation rate 0.6 units/s); steep deaccentuation, final RD value = 0.9 (rate 1.3 units/s). These changes in RD resulted in following changes is EE: no deaccentuation, final EE value = 67.2 dB; shallow deaccentuation, final EE value = 65.6 dB (5.1 dB/s); steep, final value = 64.3 dB (9.4 dB/s).

Peak magnitude, peak alignment and deaccentuation were manipulated individually and in combinations. The combinations of parameters are shown in Table 1. Overall, 20 combinations were synthesized for each of the two syllables WAY and YEAR. The total number of stimuli used in the listening test was 41 (2 syllables x 20 combinations + 1 baseline stimulus). It should be noted that in perceptual terms, differences in peak magnitude and peak timing are registered in terms of rate of signal change.

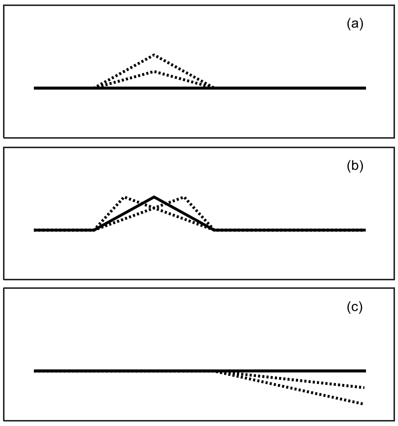

Figure 1: *Schematic of parameter manipulation in the synthesized stimuli: (a) peak height; (b) peak alignment (rate of change); (c) post-focal deaccentuation.*

3. Listening tests

In the listening tests, which were conducted online, the 41 synthesized stimuli were presented to the participants in random order. The participants were advised to use high quality headphones during the test. The listeners were informed that they were going to hear a number of utterances in which the syllables WAY or YEAR may or may not be realized as prominent. The participants were asked to listen to each stimulus as many times as they wish and to complete a number of tasks. Two listening tests were carried out. In the first, the participants' tasks were as follows:

1) Select the prominent syllable (WAY, YEAR, Neither);

2) For the prominent syllable, indicate the magnitude of prominence, using a slider on a continuous analogue visual 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').

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 analogue visual scale. They were also asked to rate the naturalness of the stimuli and to indicate how confident they were in their judgment on a continuous analogue visual scale.

The first experiment was completed by 29 participants; the second experiment was done by 18 participants.

4. Results

Listening Test 1. 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, and that the degree of prominence perceived on the targeted syllable would correlate with the magnitude of the source manipulation carried out. The results show a clear difference in how the two syllables were rated. The overall confusion matrix is given in Table 3. In the majority of cases, the WAY-manipulated sentences (those in which the WAY syllable and following material were manipulated) were identified as having prominence on WAY (64%). For the YEAR sentences (those where the YEAR syllable and subsequent material were similarly manipulated), listeners were as likely to hear prominence on WAY as on YEAR – in other words, these sentences were heard to be much the same as the baseline stimulus.

Table 2. *Parameter combinations in the resynthesis.*

	N	Peak magn.	Peak align.	Deaccent.
Baseline	0	0	0	0
Peak	1	low	0	0
	2	high	0	0
Peak + align.	3	low	early	0
	4	low	late	0
	5	high	early	0
	6	high	late	0
Deaccent.	7	0	0	shallow
	8	0	0	steep
Peak + deaccent.	9	low	0	shallow
	10	low	0	steep
	11	high	0	shallow
	12	high	0	steep
Peak + align. + deaccent.	13	low	early	shallow
	14	low	late	shallow
	15	high	early	shallow
	16	high	late	shallow
	17	low	early	steep
	18	low	late	steep
	19	high	early	steep
	20	high	late	steep

Table 3. *Overall confusion matrix of perception of the stimuli in the listening test.*

		Modified sentences and baseline		
		WAY	YEAR	BASELINE
Perceived	WAY	64%	37%	38%
	YEAR	19%	39%	38%
	Neither	17%	23%	24%

The results of Test 1 concerning the identification of the syllables as prominent for the individual stimuli are shown in Figure 2, and these are listed in the order given in Table 2. Responses to the baseline stimulus are also shown as the leftmost bar in each panel. Clearly, more stimuli were selected as having a prominent WAY syllable in the WAY sentences than YEAR in the YEAR sentences. In the former case (WAY sentences), the stimuli for which 70% or more of the listeners agreed in prominence identification on WAY included mainly those with high peaks and steep postfocal deaccentuation. Conversely, the stimuli which entailed deaccentuation alone, or manipulations involving a low peak, were identified as prominent by fewer participants. For the stimuli containing manipulations to the YEAR syllable and following material, results were very different: here, there were only relatively minor shifts from the baseline stimulus results.

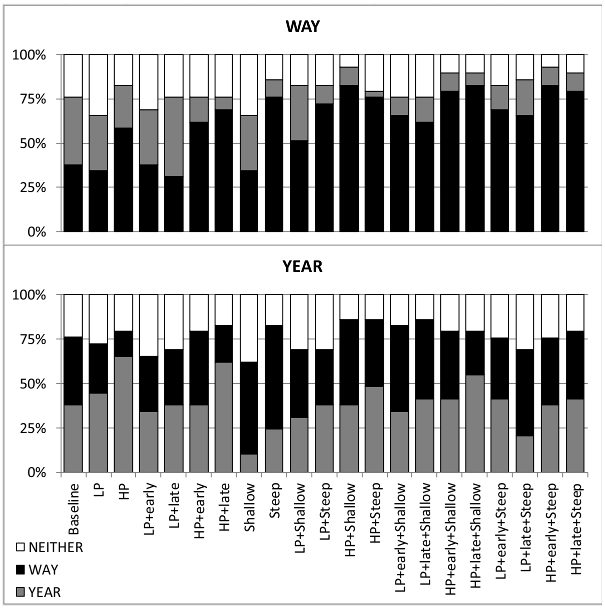

Figure 2: *Frequencies (%) with which WAY, YEAR and Neither selected as prominent.*

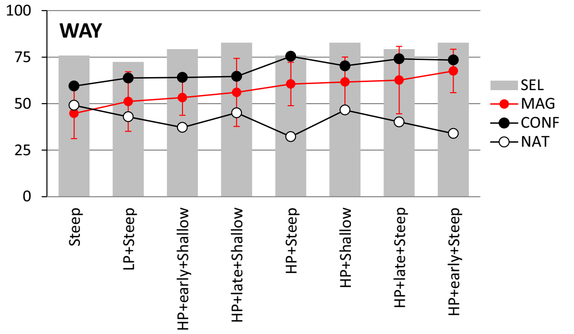

Figure 3: *Prominence magnitude, confidence and naturalness for cases where WAY deemed prominent by 70% or more.*

Figure 3 shows the mean ratings and 95% confidence intervals of prominence magnitude as well as confidence and naturalness ratings for those cases where the syllable WAY was deemed prominent by 70% or more of listeners (grey bars). The peak height appears to be the main determinant of perceived prominence. There is a significant positive correlation between listeners' judgments on the degree of prominence and their confidence in such judgments ($r = 0.91$, $n = 8$, $p < 0.05$) and a less consistent negative correlation with naturalness ($r = -0.59$, $n = 8$, $p = 0.06$).

A 3 x 3 (peak height, peak alignment, deaccentuation slope) factorial analysis was conducted to establish the contribution of the type of parameter manipulation to the prominence magnitude rating. Results indicated a significant main effect of peak F(2,583) = 15.31, p < 0.01 and deaccentuation F(2,583) = 10.35, p < 0.01. There was also a weak but significant interaction effect of peak and deaccentuation F(4,583) = 2.45, p < 0.046. The effect of peak alignment was not significant F(2,583) = 1.46, p = 0.23.

4.1. Listening Test 2

In this test, participants marked the relative prominence of all the syllables in the utterance. Figure 4 illustrates results, in terms of the difference in the perceived magnitude of the WAY and YEAR syllables within each of the stimulus sentences. (Positive values = WAY perceived as more prominent; negative values = YEAR perceived as more prominent. Blue and red bars indicate sentences with manipulations that should in principle enhance prominence of WAY and YEAR respectively. The cases where the difference in the magnitude of perceived prominence of WAY and YEAR is significant are shown by asterisks.)

The results here again show a clear difference in how prominence is rated in the two cases. In the case of the WAY sentences, most manipulations did enhance the relative prominence of the WAY syllable. The most striking (and statistically significant) effects are found when there is both a high peak on the syllable WAY, alongside deaccentuation of the postfocal material. The steepness of the deaccentuation (shallow or steep) in these cases does not appear to matter. Where the peak on WAY is lower, the effects are less, and only achieve significance when the low peak combines with deaccentuation, steep or shallow. Manipulation to height of the WAY peak, on its own, does increase that syllable's relative prominence, but this increase only renders it significantly different in prominence from YEAR when the peak is high and it is aligned to be early or late. Manipulating the deaccentuation on its own (without adjusting the WAY peak height) is effective only when a steep deaccentuation slope is used.

Ratings for the sentences where manipulations should in principle lead to enhancing the prominence of YEAR are very different (red bars). Although a few stimuli shifted the balance somewhat (e.g., some cases where YEAR had a high peak) there was not a single case where such a shift was significant. In most cases the relative prominence of the two syllables was rather similar to the baseline stimulus.

5. Discussion

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 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 and deaccentuation in the postfocal material.

In the final accentable syllable (YEAR) the findings were not symmetrical. A low peak has a negligible effect: a high peak can raise the perceived prominence, but the effects are not significant. Furthermore, postfocal source deaccentuation does not appear to play a role. The lack of a deaccentuation effect here may simply reflect the fact that there are only two unstressed syllables for deaccentuation to play out, and that this is insufficient, and not comparable to the case of WAY.

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 F0. The F0 was kept constant in these stimuli as the objective was to ascertain the role of other voice source effects. However, F0 movement co-occurs with the kinds of source effects implemented here and it is very likely that F0 movement is far more crucial in final than in non-final syllables. In a production study of focus [19] an F0 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 [1] indicated that while accented syllables in non-final position may, but need not, exhibit F0 movement, a sharp F0 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.

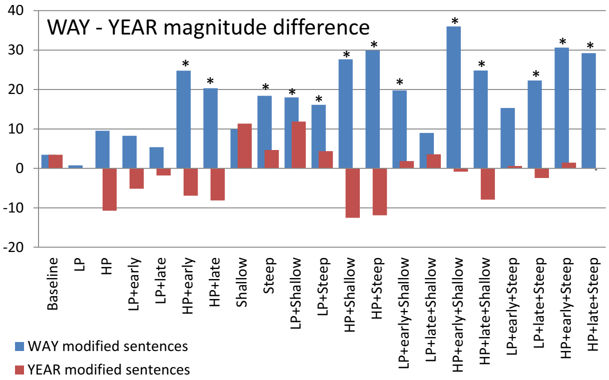

Figure 4: *The WAY-YEAR difference in magnitude of perceived prominence. * = significant difference between WAY and YEAR in the same utterance.*

6. Conclusions

These tests indicate that voice source modulations of the type observed in production data can cue focal prominence. They further suggest that having a source prominence peak on the focally accented syllable may work synergistically with a degree of source deaccentuation in the postfocal material.

It was striking however that the manipulations that induced the perception of focal accentuation in the non-final syllable had much less effect on the final syllable, where on the whole, 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 F0 movement (not included) is a necessary component. As a next step in these studies, we hope to look at the interplay of source parameters with F0 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 in focally accented syllables.

7. Acknowledgements

This research is supported by ABAIR (An Roinn Ealaíon, Oidhreachta & Gaeltachta) and CabairE (An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta). We thank Christopher La Pat for assistance with the setup of the listening test.

8. References

[1] A. Ní Chasaide, I. Yanushevskaya, J. Kane, and C. Gobl, "The Voice Prominence Hypothesis: the interplay of F0 and voice source features in accentuation," presented at the Interspeech 2013, Lyon, France, 2013.

[2] I. Yanushevskaya, C. Gobl, J. Kane, and A. Ní Chasaide, "An exploration of voice source correlates of focus," presented at the Interspeech 2010, Makuhari, Japan, 2010.

[3] A. Ní Chasaide, I. Yanushevskaya, and C. Gobl, "Prosody of voice: declination, sentence mode and interaction with prominence," presented at the XVIIIth International Congress of Phonetic Sciences, Glasgow, UK, 2015.

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

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

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

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

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

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

[10] C. Gobl, I. Yanushevskaya, and A. Ní Chasaide, "The relationship between voice source parameters and the Maxima Dispersion Quotient (MDQ)," presented at the Interspeech 2015, Dresden, Germany, 2015.

[11] A. Ní Chasaide, C. Gobl, and P. Monahan, "A technique for analysing voice quality in pathological and normal speech," *Journal of Clinical Speech and Language Studies,* vol. 2, pp. 1-16 1992.

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

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

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

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

[16] I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "The interaction of long-term voice quality with the realisation of focus," presented at the Speech Prosody 2016, Boston, MA, forthcoming 2016.

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

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

[19] A. Ní Chasaide, I. Yanushevskaya, and C. Gobl, "Voice source dynamics in intonation," presented at the XVIIth International Congress of Phonetic Sciences, Hong Kong, China, 2011.

[20] C. Gobl, "Voice source dynamics in connected speech," *STL-QPSR,* vol. 1, pp. 123-159, 1988.