INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

R_d as a control parameter to explore affective correlates of the tense-lax continuum

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

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

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

Abstract

This study uses the R_d glottal waveshape parameter to simulate the phonatory tense-lax continuum and to explore its affective correlates in terms of activation and valence. Based on a natural utterance which was inverse filtered and source-parameterised, a range of synthesized stimuli varying along the tense-lax continuum were generated using R_d as a control parameter. Two additional stimuli were included, which were versions of the most lax stimuli with additional creak (lax-creaky voice). In a listening test, participants chose an emotion from a set of affective labels and indicated its perceived strength. They also indicated the naturalness of the stimulus and their confidence in their judgment. Results showed that stimuli at the tense end of the range were most frequently associated with *angry*, at the lax end of the range the association was with *sad*, and in the intermediate range, the association was with *content*. Results also indicate, as was found in our earlier work, that a particular stimulus can be associated with more than one affect. Overall these results show that R_d can be used as a single control parameter to generate variation along the tense-lax continuum of phonation.

Index Terms: voice source, affect, speech synthesis, R_d, perception

1. Introduction

The expression of emotion is one of the most difficult aspects of natural speech to replicate with speech synthesizers. For many years, the main way in which human feeling was added to synthetic speech was through pitch, amplitude and speech rate manipulations. Other methods involved complicated speech parameter manipulations that required expert knowledge [1, 2], using extensive emotional speech corpora [3, 4], or utilizing speaker adaptation and transformation methods [5]. The resources required for these methods are not available for many languages, which require an alternative method for producing expressive speech synthesis.

It is generally accepted that voice quality plays an important role in signaling speaker affect [6-9]. Control of the voice source is therefore an important aspect in building expressive speech synthesis systems. [10] developed a method that allows for control and transformation of the voice source in statistical parametric speech synthesis. This method still requires a level of expert knowledge to carry out any transformations. Prior work [6, 11-13] showed that synthetic stimuli of basic voice quality types generated using the KLSYNN88 synthesizer [14] were consistently associated by listeners with particular affective states. However, they involved complex parameter manipulations that may not necessarily be suited to real-time speech technology applications. The optimal solution would be to implement a control system that requires changing a minimal set of voice source parameters that translates to a change in the perceived affect of synthesized speech. This would have applications in developing realistic personalized and expressive voices, generating characters for educational games and other software, such as screen readers and spoken dialogue systems for low resource languages [15, 16].

In this study, we attempt to implement an element of such a control system using the glottal waveshape parameter, R_d [17, 18]. The R_d parameter offers a way to reduce the number of voice source parameters required in controlling the perceived affect of synthetic speech. A previous study has already found that this parameter is effective at signalling focal prominence in the absence of f_0 variation [19].

This study involved generating a set of synthetic stimuli varying along the tense-lax continuum using the R_d parameter. The association of these stimuli with a number of affective states was investigated in a listening test.

2. Materials and Method

2.1. Synthetic stimuli

The stimuli were constructed on the basis of an all-voiced utterance ‘We were away a year ago’ spoken by a male speaker of Irish English. The utterance was produced with narrow focus on the WAY syllable. It was originally recorded for another study, where other versions of the sentence, with differing focal placement, were also obtained and their source parameters analyzed [20]. The utterance was inverse filtered using interactive manual inverse filtering software [21]. Parameterization of the voice source was then carried out using the Liljencrants-Fant (LF) model [22], on a pulse-by-pulse basis. Based on the analyzed utterance, a modal voice stimulus was first generated (see Section 2.1.2). The synthetic stimuli, varying along the tense-lax continuum, as well as lax-creaky stimuli, were then produced using a range of values for the global waveshape parameter R_d (see Sections 2.1.3 and 2.1.4).

2.1.1. The R_d parameter

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

$$R_d = \frac{1}{0.11}\left(f_0 \times \frac{U_p}{E_e}\right) \tag{1}$$

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

http://dx.doi.org/10.21437/Interspeech.2017-1448

the glottal pulse. Note that U_p/E_e 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 R_d equal to the declination time in milliseconds when f_0 is 110 Hz [17].

Variation in R_d tends to reflect voice source variation along the tense-lax continuum; the values typically range between 0.5 (tense voice) to 2.5 (lax voice), with modal voice having a value of approximately 1. Changes to R_d cause other LF parameters such as R_a and R_k to vary. These changes can be predicted from R_d. (A full description of the various glottal parameters can be found in [23]).

In this study, R_d was manipulated within the range observed for the speaker analyzed (0.49 – 2.84). In order to generate the LF model glottal waveform, a full set of LF model parameters are needed. These were obtained from R_d using the parameter correlations presented in [17].

2.1.2. Modal stimulus

A synthetic stimulus for modal voice was generated using parameter values obtained from the prior inverse filtering and source parameterization. The resulting stimulus sounded rather tense; the average parameter values for this stimulus (f_0 = 120 Hz, R_d = 0.8, E_e = 71 dB) also corresponded to a tense phonation type. To make the voice laxer to improve its naturalness the R_d parameter values were increased by 25% to a mean value of 1.0. These parameters were then used to recalculate E_e, resulting in an average value of 63 dB. The contour of the original utterance was stylized by taking values of f_0, E_e and R_d at vowel midpoints in each syllable and interpolating between them. A short transitional period was also added at the beginning and end of the contour to capture the onset and offset of phonation. This made the parameter contours easier to process for subsequent stimuli generation. The resulting contour for the R_d parameter is shown in Figure 1 (black line). The stimulus was synthesized using cascade formant synthesis.

2.1.3. Lax and tense stimuli

Once the modal stimulus had been created, its parameter values were used as the basis for synthesizing the remaining stimuli. Due to the greater range between the modal and upper R_d limit (1 to 2.84) compared to the range between the modal and lower R_d limit (1 to 0.49), a logarithmic scale was used to establish four steps that would be approximately equidistant in the direction of both tense and lax voice. The steps are shown in Figure 1.

For all the stimuli, f_0 values were kept the same as those of the modal stimulus. E_e was recalculated according to the new R_d values for each stimulus. The stimuli were labelled as tense1, tense2, tense3, tense4, lax1, lax2, lax3, and lax4, where tense4 and lax4 had the lowest and highest values of R_d respectively. To improve the naturalness of the lax stimuli aspiration noise was added to the source waveform using the method described in [24]. This method automatically determines the overall amplitude and modulation of the added aspiration noise based on the shape of LF model waveform.

2.1.4. Lax-creaky stimuli

Two further stimuli were generated by adding aperiodicity to the source signal of lax3 and lax4. This effectively mimicked a lax-creaky voice quality. This was achieved by using a method similar to the one with which diplophonic double pulsing is produced in the KLSYN88 synthesizer [14]. These lax-creaky stimuli were included here based on the findings of a previous study [6] where this voice quality was found to yield high ratings for low activation affective states such as bored, intimate, relaxed, and content. The addition of creak to an already lax synthetic stimulus would not require much more in terms of parameter input, but could add an extra dimension of affective control.

To generate creak, the source signal was divided into frames of two pulses each. The first pulse of each frame was shifted towards the second pulse and attenuated by an equal percentage. This percentage was determined by normalizing and scaling the f_0 contour of the utterance to the range of 0 to 10%. These values were then inverted. This corresponded to a 10% shift and attenuation at the points in the synthetic stimuli with the lowest f_0 values, and no shift or attenuation at the points with the highest f_0. This method, along with a reduction in aspiration noise, produced satisfactory lax-creaky stimuli, labeled as lax3+c and lax4+c.

The total number of stimuli generated was 11 (modal, 4 x tense, 4 x lax, 2 x lax-creaky).

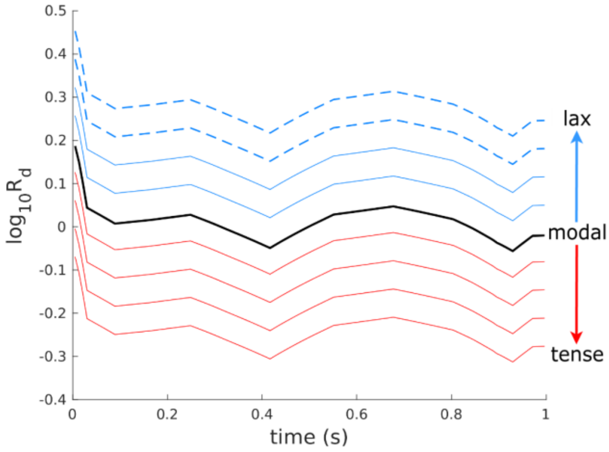

Figure 1: *$\log_{10} R_d$ values of the synthetic stimuli. The lax and tense stimuli are shown in blue and red respectively. The modal stimulus is shown in black. The values used for the two lax-creaky stimuli are indicated by the blue dashed lines.*

2.2. Listening test

The 11 synthetic stimuli, as well as the original natural utterance, were presented to 30 participants (all native speakers of English) in a listening test. The listening test was carried out in a quiet environment using high quality headphones, via an interactive GUI. The participants were presented with 60 synthesized stimuli (5 repetitions of the 12 stimuli) in random order. An additional 5 random stimuli were added at the beginning of the test so that the participants could become accustomed to the process involved. The results of these first 5 stimuli were discarded. The participants were informed that they were going to hear a number of different sound files, and that they could listen to each file as many times as they wished in order to answer the following questions for each stimulus:

1) How does the speaker sound? [the subject chose from a selection of radio buttons with affective labels];

2) To what extent? [a continuous analogue visual scale ranging from 'Not at all' to 'A lot'];
3) How confident are you in your judgment? [a continuous analogue visual scale ranging from 'Not at all confident' to 'Very confident'];
4) How natural does the audio sound? [a continuous analogue visual scale ranging from 'Not at all natural' to 'Very natural'].

Participants were given eight affective labels to choose from for question 1. These were: *relaxed, angry, content, upset, happy, sad, excited* and *bored*. These emotional states were chosen to give a balanced set in terms of high and low activation, and positive and negative valence emotional dimensions (see Figure 2). Participants were also given the options of *other* and *no emotion*. The continuous analogue scale [25] used to measure the magnitude of the affective coloring present as well as the listener's confidence and the naturalness of the stimuli (questions 2-4) was interpreted as ranging from 1 to 100. The confidence score covered questions 1 and 2.

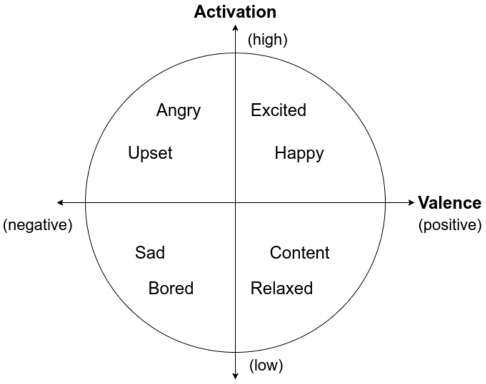

Figure 2: *Affective labels used in the listening test in the two-dimensional model of affect.*

There were no major expectations that a particular stimulus would be associated with a particular affect (prior studies showed that the same voice quality can cue a number of different affects, e.g., [6, 13]), nor did we expect clear cut categorical differentiation of the selected affects. We expected support to earlier findings [12] that tenser voice quality would tend to be associated with high activation states and laxer and creakier voice quality with low activation states. Although positive high activation states (*happy, excited*) were included to ensure a balanced representation, we did not expect to achieve high percentage of stimulus to affect association here. The type of manipulations used in the construction of the stimuli did not include large dynamic variation of f_0, nor did it include any adjustments in the filter settings. These, however, are well known to be necessary to generate the impression of excitement and happiness (shorter vocal tract, higher frequency of the second formant for smiling speech).

3. Results and Discussion

Table 1 shows the percentage of cases in which a particular stimulus was associated with each of the options in the listening test. Overall, tense voices (except tense1) as well as the natural production were mainly associated with the *angry* affect, lax-creaky and lax voices (except lax1) were associated with the *sad* affect. Modal voice as well as the least extreme cases of lax and tense voices (lax1 and tense1) were associated with *content*.

Table 1: *Cases (%) in which the stimuli were associated with the affects in the listening test. The most frequent case for each stimulus is underlined and emboldened.*

	Angry	**Upset**	**Excited**	**Happy**	**Content**	**Relaxed**	**Bored**	**Sad**	**Other**	**No emotion**
natural	**31**	7	8	12	17	7	1	5	3	9
tense4	**27**	6	21	11	13	5	1	3	3	12
tense3	**26**	8	12	9	16	6	3	1	2	17
tense2	**21**	10	10	10	17	4	5	5	2	16
tense1	15	12	4	6	**21**	9	5	4	2	21
modal	14	10	8	9	**23**	6	5	1	4	19
lax1	2	16	3	3	**21**	7	16	13	1	17
lax2	3	16	1	1	7	15	19	**22**	2	13
lax3	5	21	1	0	5	12	18	**23**	2	13
lax4	3	20	1	1	5	7	19	**34**	1	8
lax3+c	5	15	0	1	3	4	17	**30**	2	22
lax4+c	7	14	1	3	1	8	13	**29**	3	22

Looking at the two most frequently selected affects (Figure 3), a clear trend can be observed. The degree of tenseness/laxness in the synthetic stimuli is correlated with the frequency with which a particular stimulus is associated with anger/sadness. As in earlier studies, no one-to-one mapping was found between voice quality and affect. For example, the most frequent response for the lax stimuli was sad, but bored and upset were also often chosen.

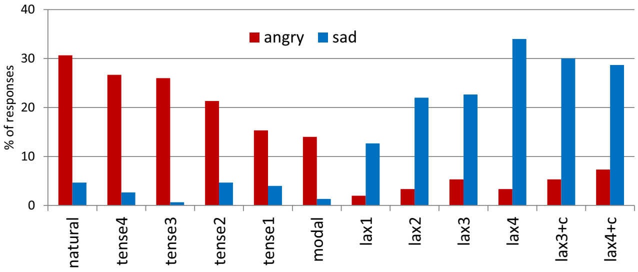

Figure 3. *Frequency of responses associating the stimuli with the angry and sad affects.*

Due to the wide distribution of selected affects per synthetic stimuli, we further grouped affects with similar valence and activation (see Figure 2). The results are plotted in Figure 4.

Tense voice was most commonly associated with high activation affects and lax-creaky and lax voices – with low activation affects. Modal voice was almost equally associated with high and low activation affects.

The natural utterance was more commonly associated with high activation states. This is most likely caused by the fact that it has a lower R_d value (between tense1 and tense2) than the modal synthetic stimulus.

In terms of valence, the lax voice stimuli were mainly associated with negative affects, whereas tense, natural and modal stimuli were equally associated with both negative and positive affects.

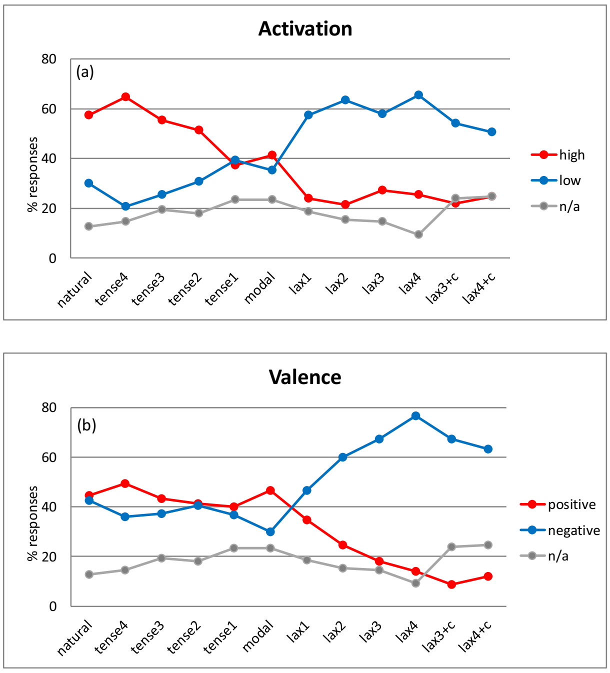

Figure 4. *Percentage of instances where the synthetic stimuli were associated with affects of particular activation (panel a) or valence (panel b) group.*

This even distribution agrees with the initial predictions that additional manipulations of the vocal tract are necessary to differentiate between positive and negative valence states for stimuli representing tenser phonation types.

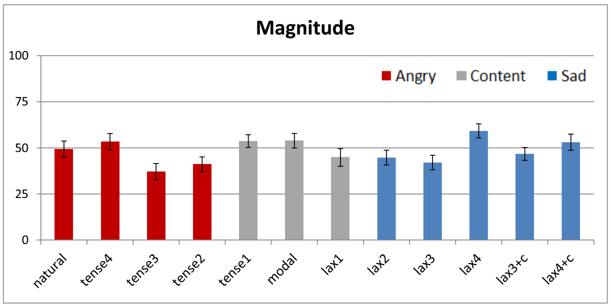

Figure 5. *Magnitude of most frequently selected affects for each stimulus (mean and standard error).*

The magnitude of the manipulation correlated with the magnitude of perceived affect, particularly with the stimuli at the limits of the tense-lax range. The addition of creak did not yield consistent results: it seemed to increase the magnitude slightly in the case of lax3, but reduced the magnitude in lax4.

The modal stimulus was associated with a different affect thn the natural utterance. This is most likely due to the fact that the source parameter values used to generate the modal stimulus were altered to produce a laxer voice quality. The natural stimulus had parameter values between the tense1 and tense2 stimuli. It is not surprising that these stimuli were associated with affects of the same valence/activation, and that the magnitude of the perceived affects was also similar for these stimuli. The small discrepancy between them is perhaps due to artifacts introduced by the analysis and synthesis process.

The mean naturalness score for each stimulus can be seen in Table 2. The natural stimulus was found to have the highest naturalness at 76. Modal, tense1, lax1, and lax2 follow this with values between 61 and 63. As the stimuli become more lax or tense, there is a slight drop in naturalness, but the values are still well above 50. The two lax-creaky stimuli show the lowest naturalness score. This is likely due to the values used in the generation of these stimuli which may not have been optimal despite the fact that creak was varied dynamically. More research is needed to improve this feature.

All the mean confidence score values were above 54, with the natural stimulus having the highest confidence score of 65.

Table 2: Mean naturalness and confidence score for each stimulus (scale range 0-100).

	Naturalness	Confidence
natural	76	65
tense4	57	56
tense3	57	57
tense2	59	56
tense1	63	57
modal	62	54
lax1	62	55
lax2	61	55
lax3	58	56
lax4	58	61
lax3+c	42	58
lax4+c	43	58

4. Conclusions

The study showed that stimuli ranging in the tense-lax continuum generated by manipulating a global waveshape parameter R_d can evoke an affective response and that the direction of the response correlates with the magnitude of the manipulation. Thus, the tenser the voice, the higher the perceived magnitude of, say, anger.

Although manipulations of R_d were effective in evoking an affective response in terms of activation, they were not so successful in distinguishing between valence states. This is most likely due to the lack of vocal tract parameter manipulations. Future work will explore control of these parameters. It will also include improving the quality of the synthetic creak.

Overall these results show that R_d can be used as a single control parameter to generate variation along the tense-lax continuum of phonation. This has potential uses in the development of expressive speech technology applications.

5. Acknowledgements

This research is supported by the Dept. of Arts, Heritage, Regional, Rural and Gaeltacht Affairs (*ABAIR* project) and CabairE (An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta). We thank Christopher La Pat for assistance with the setup of the listening test.

6. References

[1] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion," *Journal of the Acoustical Society of America,* vol. 93, pp. 1907–1108, 1993.

[2] J. Cahn, "The generation of affect in synthesised speech," *Journal of American Voice I/O Society,* vol. 8, pp. 1-19, 1990.

[3] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generations of databases," *Speech Communication,* vol. 40, pp. 33-60, 2003.

[4] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Communication,* vol. 40, pp. 161-187, 2003.

[5] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, "Constructing emotional speech synthesisers with limited speech database," in *Interspeech 2004*, Jeju Island, Korea, 2004, pp. 1185-1188.

[6] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication,* vol. 40, pp. 189-212, 2003.

[7] K. R. Scherer, "Vocal communication of emotion: a review of research paradigms," *Speech Communication,* vol. 40, pp. 227-256, 2003.

[8] S. Patel, K. R. Scherer, E. Björkner, and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biological Psychology,* vol. 87, pp. 93-98, 2011.

[9] J. Sundberg, S. Patel, E. Björkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Transactions on Affective Computing,* vol. 2, pp. 162-174, 2011.

[10] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4704-4707.

[11] C. Gobl, E. Bennett, and A. Ní Chasaide, "Expressive synthesis: how crucial is voice quality?," in *IEEE Workshop on Speech Synthesis*, Santa Monica, California, USA, 2002, pp. 1-4.

[12] C. Ryan, A. Ní Chasaide, and C. Gobl, "Voice quality variation and the perception of affect: continuous or categorical?," in *XVth International Congress of Phonetic Sciences*, Barcelona, Spain, 2003, pp. 2409-2412.

[13] I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Universal and language-specific perception of affect from voice," in *XVIIth International Congress of Phonetic Sciences*, Hong Kong, China, 2011, pp. 2208-2211.

[14] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America,* vol. 87, pp. 820-857, 1990.

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

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

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

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

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

[20] C. Gobl, I. Yanushevskaya, and A. Ní Chasaide, "The relationship between voice source parameters and the Maxima Dispersion Quotient (MDQ)," in *Interspeech 2015*, Dresden, Germany, 2015, pp. 2337-2341.

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

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

[23] 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.

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

[25] D. L. Streiner and G. R. Norman, *Health Measurement Scales*, 4 ed. Oxford: Oxford University Press, 2008.