10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Testing the GlórCáil System in a Speaker and Affect Voice Transformation Task

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 paper describes the results of a voice transformation task experiment conducted as part of the evaluation of a speech synthesis system (the GlórCáil system, also described). The participants were required to manipulate the system's control parameters reflecting changes in voice quality, f_0 and vocal tract length of the speaker (VT) in synthetic utterances. A synthetic baseline utterance was manipulated to make it sound like a target speaker (*man, woman, child*) with affective colouring (*sad, angry, no emotion*). The control parameters of the system proved useful in modulating speaker characteristics and paralinguistic prosody. The manipulations performed by the participants were mainly in the expected direction. f_0 and VT were found to be significant predictors of speaker gender/age, but not of affect. The voice quality related parameter R_d was a significant predictor of affect, but not of speaker identity. Significant interactions of predictors were found for f_0 and VT. The control parameter values obtained in this experiment will be used to generate stimuli to test the proposed system when it is integrated into a DNN-based speech synthesis system as part of the ongoing work of the ABAIR project.

Index Terms: speech synthesis, voice quality, voice transformation, affect, speaker characteristics

1. Introduction

Flexible expressive speech synthesis is desirable in many applications. This paper describes a system that allows for the transformation of voices by using our understanding of the speech production process, and in particular, by using an acoustic glottal model-based excitation that more closely models the natural glottal source. Although many state-of-the-art methods have implemented improved excitation models that increase the naturalness of synthetic speech [1] [2] [3], they are complicated to implement and control for non-expert users.

The ABAIR Irish synthesis project [4, 5], to which this work is linked, requires flexible, expressive voices in applications for users with disabilities, such as visual impairment, as well as in Irish language learning applications, such as interactive multimodal educational games. Flexible and easy control of the voice source and vocal tract filter characteristics of synthetic speech would be a major benefit in producing expressive material for these games and allow users with communication disorders to express themselves through a synthetic voice. The interactive educational games [4] entail scenarios with many characters. A limited number of Irish voices have been developed, and so a limited number of characters are available to choose from. These characters need expressive voices that match the contexts of the scenarios within the games.

Populating these educational games with characters who have different voices has, to date, been achieved using relatively crude transforms. These voices cannot achieve the diversity and natural quality of human voices. The ideal would be a system that allows for control of both the voice source and the vocal tract filter, thus allowing for the generation of, from first principles of speech production, any number of voices differing from each other in subtle or major ways.

It would be desirable also that the required transformations be achieved by a small set of control parameters. Our earlier studies explored control of linguistic and paralinguistic prosody in synthesis using a small set of parameters and focused mainly on the global waveshape parameter R_d [6, 7]. The findings suggest that changes in voice source parameters can lead to changes in the perceived prominence of words/syllables [8], [9] as well as changes in the perceived affective colouring of an utterance [10]. R_d was found important in describing cross-speaker differences in voice quality [11] and was implemented in synthesis systems also in [12] [1] [13]. Overall, the results of these studies suggest that R_d is an effective voice quality control parameter. By reducing the dimension of source parameters to a minimal set, the process of parameter manipulation is simplified, which is important when considering usability and integration into statistical parametric speech synthesis (SPSS) systems.

The considerations outlined above lead to the development of the GlórCáil analysis and synthesis system. The name comes from the Irish words for voice (glór [ˈɡl̪ˠoːɾʲ]) and quality (cáilíocht [ˈkɑːlʲiːxt̪ˠ]). This paper provides a brief description of the system and describes the results a voice transformation task that evaluates the system's ability to alter such speaker characteristics as gender, age and affective colouring of speech.

2. System description

The GlórCáil system is an application developed for the analysis and resynthesis of speech with a particular focus on control of voice source parameters. It is implemented in the MATLAB environment [14]. Vocal tract and voice source parameters are first estimated during the analysis stage. Speech is then resynthesized using an acoustic glottal source model in place of the original glottal source. This allows users to manipulate voice source parameter contours in an interactive GUI, listen to the results of manipulations and make further desired changes to the perceived voice quality of the utterance.

2.1. Analysis stage

First, f_0, GCI locations and voiced/unvoiced regions are estimated in an audio recorded signal using the REAPER program [15]. Next, inverse filtering is carried out on a frame-by-frame basis using the modified version of Iterative Adaptive Inverse Filtering [16], GFM-IAIF [17], with a frame length of

10.21437/SpeechProsody.2020-194

25 ms and a frameshift of 5 ms. Discrete all pole (DAP) modelling is used in place of LPC modelling in the GFM-IAIF process as it provides more accurate estimations of the vocal tract transfer function [18]. This analysis provides estimations of the differentiated glottal flow and the filter coefficients describing the vocal tract transfer function. The filter coefficients are converted to line spectral frequencies (LSFs) to make them less susceptible to distortions introduced by later processing steps and to make them more robust to statistical modelling in the cases where the vocoder is used in conjunction with SPSS system. Source parameterisation is then performed on each pulse using a method based on dynamic programming [19].

2.2. Synthesis stage

Once parameters have been extracted from an utterance, the GlórCáil system can be used to resynthesize it. First, the voiced and unvoiced excitations are generated. Voiced regions are defined by frames with f_0 values above and below a minimum and maximum frequency (default 50 and 500 Hz respectively). The voiced excitation consists of Liljencrants-Fant (LF) model pulses [6, 20] with the addition of amplitude modulated Gaussian white noise. The f_0 value from the first voiced frame is used to calculate the period of the first LF pulse using the relationship $f_0 = 1/T_0$. The period is used to define the start and end points of the pulse frame within the voiced excitation.

The parameters E_e, R_d and f_0 are used to generate the full set of LF model [20] parameters using the correlations outlined in [6], which are in turn used to generate the corresponding LF pulse. The R_d parameter is derived from f_0, E_e and U_p as follows: $(1/0.11)\times(f_0 \cdot U_p/E_e)$, where E_e is the excitation strength (measured as the negative amplitude of the differentiated glottal flow at the time point of maximum waveform discontinuity) and U_p is the peak flow of the glottal pulse (see Figure 1). Variation in R_d is proposed to reflect voice source variation along the tense-lax continuum; the values typically range between 0.3 (tense voice) to 2.5 (breathy voice).

Amplitude modulated noise is added to the pulse according to the method described in [21]. The next pulse is calculated using parameter values from the closest frame to the end of the previous pulse. This process continues until the current voiced region has been filled with pulses. The pulses start and finish at zero, so no windowing is required to prevent discontinuities in the signal. Frames with values below the minimum f_0 threshold are treated as unvoiced regions; white Gaussian noise is used as the excitation signal in these areas.

The generated excitation signals are filtered by filters that represent the vocal tract transfer function. The voiced excitation is filtered using the filter coefficients obtained from glottal inverse filtering. A scaling factor can be applied to the filter coefficients to effectively shorten or lengthen the vocal tract. The frequencies of the poles are warped by performing a bilinear transformation in the z-domain, based on the implementation in [22]. This approach involves changing the original filter coefficients through the substitution of z^{-1} (unit delay filter) with an all-pass filter, as shown in Equation (1) [23]

$$z^{-1} \rightarrow \tilde{z}^{-1} = \frac{z^{-1}-\lambda}{1-\lambda z^{-1}} \quad (1)$$

where λ is the warping factor. The range of the warping factor is limited to between -0.1 and 0.1, where negative and positive values effectively shorten or lengthen the vocal tract respectively. This ability to modify the resonances of the vocal tract adds an extra dimension of control and allows for further transformations to be made to the synthetic speech. The unvoiced excitation is filtered using filter coefficients and gain values obtained from the regular LPC analysis. The final speech signal is then created by overlap-adding each of the filtered frames.

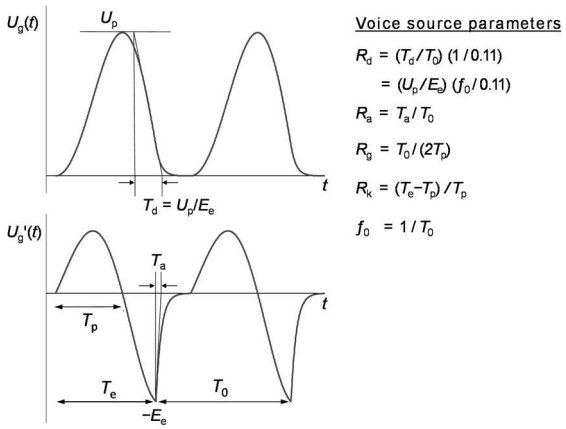

Figure 1: *Parameters used to generate the LF model waveform (adapted from [6]). Upper panel: glottal flow; lower panel: glottal flow derivative.*

An earlier study [24] tested the functionalities of the described system in a manipulation task in which user-driven data were obtained of perceptually salient contours of R_d in signalling focal prominence. This paper describes the abilities of the GlórCáil system to transform the paralinguistic prosody of speech as well as speaker characteristics. This was tested in a manipulation task where participants were asked to modify a set of parameters in order to transform a baseline utterance to sound like a target speaker (e.g., male, female) expressing a particular affect (e.g., angry, sad).

3. Material and method

3.1. Baseline stimulus

The baseline stimulus for the voice transformation task was based on a recording of an all-voiced sentence 'We were away a day ago' spoken by a male Irish English speaker. The utterance was analysed and parameterised using the GlórCáil system analysis stage. The extracted f_0 and R_d parameter mean values were 104 Hz and 0.94 respectively. The extracted f_0 and R_d parameter contours were used in the resynthesis of the baseline stimulus using the GlórCáil synthesis stage (see Section 2.2). This baseline stimulus was subsequently modified by the participants in the listening test.

3.2. Voice transformation task user interface

A user interface was designed for the voice transformation task (see Figure 2). This interface allows users to alter the parameter contours of the baseline stimulus. Three parameters were manipulated, R_d, f_0 and the length of the vocal tract (VT). These parameters were represented in the interface by the sliding blocks labelled *Voice*, *Pitch* and *Size* respectively. The blocks, when dragged up and down, controlled parameter scaling factors. In the resynthesis of the manipulated utterance, the original (baseline) R_d and f_0 contours were multiplied by the obtained R_d and f_0 scaling factors. The scaling factor for R_d ranged between 0.5 and 2, so that the baseline values could be halved or doubled at either end of the scale. Constraints were

also applied so that the R_d values remained inside its usual range of 0.3-2.7. The scaling factor for f_0 was between 0.667 and 3, as this approximated the range of f_0 values from adult male to child (approx. 70 Hz – 300 Hz). The vocal tract scaling factor was used as a warping coefficient for modifying the vocal tract transfer function (see also Section 2.2). The vocal tract scaling factor ranged between -0.1 and 0.1, with negative values effectively simulating a shortening of the vocal tract and positive values a lengthening. Participants could listen to how their manipulations changed the utterance by pressing the *Listen* button, reset the sliders to their original starting position by pressing the *Reset sliders* button, and move on to the next page by clicking the *Next Page* button.

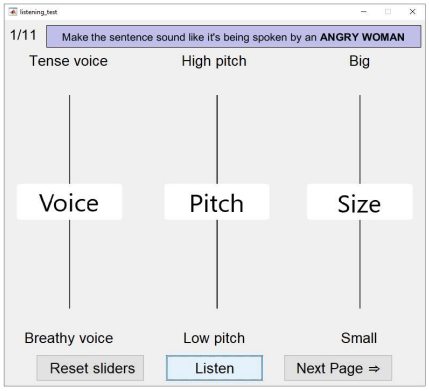

Figure 2: *User interface for voice transformation task.*

3.3. Listening test

16 participants, all native speakers of English, took part in the test. The participants were asked to manipulate the sliding blocks (and the corresponding scaling factors in the baseline sentence) so that the resulting sentence sounded like it was being spoken by a particular speaker in a certain affective state, e.g., an angry woman or a sad child. Each of the three target speakers (*man, woman, child*) was combined with three target affects (*sad, angry, no emotion*). The baseline sentence was thus to be manipulated by the participants nine times to produce nine (3 speakers x 3 affects) utterances. Two more utterances were included at the beginning of the test to allow the participants to familiarize themselves with the procedure; the results from these were discarded. The participants were allowed to listen to the results of their manipulations and make changes as many times as they wished. The stimuli were presented through high quality closed-back headphones in a quiet environment. The test took approximately 10 minutes to complete.

3.4. Hypotheses

There is a considerable body of research that has established how male, female and child speech characteristics differ in terms of voice quality, f_0 and the size of the vocal tract (e.g., [25], [26]). For example, f_0 tends to be the lowest in male speakers and the highest in children. Women have generally shorter vocal tracts than men, and children have smaller vocal tracts than women. Female talkers have been reported to have breathier phonation [27], [28] although this feature is not universal and there may be culture-specific preferences, e.g., [29]. Expression of emotions entails both laryngeal and supralaryngeal adjustments (e.g., an overall increase in the tension of the vocal tract muscles, tenser phonation and higher f_0 in the expression of anger; creaky voice, flat f_0 and lowered larynx in the expression of boredom). Expression of emotions interacts with the properties of the individual's vocal apparatus and is complicated by culture-specific pull-effects [30], [31]. Findings in [32] supported their 'size code hypothesis' suggesting that anger is encoded, among other things, by increasing the size of the vocal tract. Dynamic changes in the size of the vocal tract as well as f_0 were found in [32] to contribute to emotion identification. The goal of this current test was to confirm that the GlórCáil system parameters related to voice quality, f_0 and the size of the vocal tract can generate speech with gender/age specific characteristics with some degree of affective colouring.

Based on the descriptions of male, female and child speech characteristics in the literature, it was expected that:

- Increasingly higher f_0 scaling will be required for speaker transformation to *woman* and *child*.
- In *no emotion*, baseline-to-woman transformation would require shifts to breathier phonation (corresponding to higher R_d scaling factor values).
- In *sad* (for all target speakers), a higher R_d scaling factor value would be used, corresponding to breathier/laxer phonation.
- In *angry* (for all target speakers), a lower R_d scaling factor value would be used, corresponding to tenser phonation.
- A negative vocal tract warping factor would be used for *woman* and *child* (= shorter vocal tract).
- *Angry* voice (for all target speakers) would entail a positive vocal tract warping (= longer vocal tract).

4. Results

Mixed model analyses were performed to test if the scaling factors varied significantly across target speakers and affects. Analyses were conducted in the R environment [33] using the *lme4* [ver. 1.1-20] package [34] for model fitting. Step-down model simplification was done by eliminating non-significant effects and denominator degrees of freedom were calculated using Satterthwaite's approximation with the *lmerTest* package [35]. The models were fitted using the maximum likelihood (ML) method. The initial model included TargetSpeaker and Affect as the main predictors (fixed effects) as well as their interactions; random effects included by-subject random intercepts: Scaling ~ TargetSpeaker*Affect + (1|Participant). The model for R_d was subsequently reduced to Affect as the only fixed factor; however, we chose to keep TargetSpeaker in the final model. The models for f_0 and VT were not reduced. By-subject random intercepts were included in all models. ICC (indicative of the correlation of the items within a cluster) as well as marginal and conditional R-squared statistics [36] were obtained using the *sjPlot* package [37]. 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 estimated coefficients of the mixed effect model fitted to the scaling factors of R_d, f_0 and vocal tract (VT) obtained in the voice transformation task along with confidence intervals (CI) are given in Table 2 (see also Figure 3). Due to space constraints the tables do not show all three-way contrasts; statistically significant contrasts are reported in the text where applicable.

R_d scaling did not vary significantly across target speakers; however, significant effect of Affect was found. R_d scaling was

significantly higher in *sad* (~breathier voice) and significantly lower in *angry* (~tenser voice) for all speakers.

F0 scaling was significantly different across all speakers in *no emotion*, making it, unsurprisingly, a useful parameter for speaker differentiation (man<woman<child). The same relationship held also for *angry*. A significant TargetSpeaker*Affect interaction effect was also found for f_0 scaling. In *angry*, the f_0 scaling factor was significantly higher for a *child* target speaker than for a *woman* ($\beta = 0.24$, p=0.034). In *sad*, this differentiation was not observed. Affect was not a significant predictor of f_0 variation in *man*. In the *child* target speaker, *angry* affect required higher f_0 scaling than *sad* (β = 0.35, p=0.002); in the *woman* target speaker a similar trend was not significant.

As expected, TargetSpeaker was a significant predictor of VT warping factor values: these were significantly higher for *man* than for *woman* and *child* in all affects, and significantly larger for *woman* than for *child* (except in the *no emotion* condition). Significant TargetSpeaker*Affect interaction effect was found: the VT warping factor did not vary with affect in *woman* and *child*, but was significantly higher for the *man* target speaker in *angry* compared to *sad* ($\beta = 0.04$, p<0.001) and compared to *no emotion* ($\beta = 0.04$, p=0.001).

Table 2: *Estimated coefficients for the mixed effect models fitted to the R_d, f_0 and VT scaling factor values.*

Predictors: R_d	**β_0**	***CI***	***t***	***p***
Intercept (man)	0.1	0.92 – 1.08	25.35	<0.001
woman	0.08	0.00 – 0.16	2.02	0.043
child	0.06	-0.02 – 0.14	1.48	0.138
sad	0.60	0.52 – 0.68	14.60	**<0.001**
angry	-0.39	-0.47 – -0.31	-9.47	**<0.001**
Random effects				
ICC	0.06	Observations	144	
Marginal R^2 / Conditional R^2		0.796/ 0.808		
Predictors: f_0	**β_0**	***CI***	***t***	***p***
Intercept (man)	0.96	0.78 – 1.14	10.71	<0.001
woman	0.77	0.55 – 0.99	6.88	<0.001
child	1.08	0.87 – 1.30	9.70	<0.001
sad	0.02	-0.20 – 0.24	0.15	0.880
angry	0.07	-0.14 – 0.29	0.66	0.506
woman:sad	0.35	0.04 – 0.66	2.18	**0.029**
child:sad	0.03	-0.28 – 0.34	0.16	0.871
woman:angry	0.40	0.09 – 0.71	2.51	**0.012**
child:angry	0.32	0.01 – 0.63	2.02	**0.043**
Random effects				
ICC	0.22	Observations	144	
Marginal R^2 / Conditional R^2		0.705/ 0.770		
Predictors: VT	**β_0**	***CI***	***t***	***p***
Intercept (man)	0.02	0.01 – 0.04	2.65	0.008
woman	-0.06	-0.08 – -0.04	-5.73	<0.001
child	-0.08	-0.10 – -0.06	-7.59	<0.001
sad	-0.00	-0.02 – 0.02	-0.22	0.822
angry	0.04	0.02 – 0.06	3.40	**0.001**
woman:sad	0.02	-0.01 – -0.05	1.11	0.267
child:sad	0.00	-0.03 – 0.03	-0.15	0.883
woman:angry	-0.03	-0.06 – 0.00	-1.61	0.106
child:angry	-0.04	-0.07 – 0.01	-2.31	**0.021**
Random effects				
ICC	0.08	Observations	144	
Marginal R^2 / Conditional R^2		0.627/ 0.655		

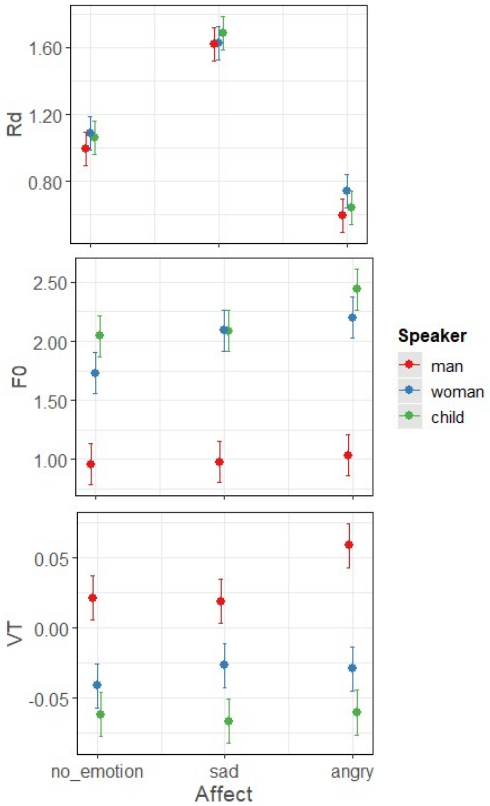

Figure 3: *Predicted values of R_d, f_0, and VT scaling factors and 95% confidence intervals.*

5. Discussion and conclusions

The goal of the test reported here was to confirm that the GlórCáil system parameter manipulations can generate speech with gender/age specific characteristics and affective colouring. The range of target speakers and affects that were specified for this voice transformation task was limited, but was deemed sufficient as a proof of concept in the initial evaluation of the system. The parameters that the system allows to control proved useful in modulating linguistic and paralinguistic prosodic characteristics in earlier studies, as well as in this study. The parameters appear to work in combination, and manipulations performed by the participants were mainly in the expected direction. As expected, f_0 and VT were found to be significant predictors of speaker gender/age, but not for affect (except in the case of VT in *angry man*). R_d was a significant predictor of affect, but not of speaker identity. Significant interactions of TargetSpeaker and Affect were found for f_0 and VT. The scaling factor values from this experiment will be used to generate stimuli to test the proposed system when it is integrated into a DNN-based speech synthesis system as part of the ongoing work of the ABAIR project.

6. Acknowledgements

This research is supported by the Government of Ireland, Department of Culture, Heritage and the Gaeltacht with National Lottery funds. This forms part of the Government's 20-year Strategy for Irish 2010-2030.

7. References

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

[2] 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.

[3] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, *HMM-based Finnish text-to-speech system utilizing glottal inverse filtering*, 2008.

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

[5] A. Ní Chasaide, N. Ní Chiaráin, H. Berthelsen, C. Wendler, and A. Murphy, "Speech technology as documentation for endangered language preservation: the case of Irish," in *XVIIIth International Congress of Phonetic Sciences*, Glasgow, Scotland, 2015.

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

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

[8] 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.

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

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

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

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

[13] S. Huber and A. Roebel, "On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system," in *16th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, Dresden, Germany, 2015.

[14] The MathWorks Inc., "MATLAB version 9.4," ed, 2018.

[15] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," ed, 2015.

[16] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication,* vol. 11, pp. 109-118, 1992/06/01/ 1992.

[17] O. Perrotin and I. V. McLoughlin, "A spectral glottal flow model for source-filter separation of speech," presented at the ICASSP, Brighton, UK, 2019.

[18] P. Alku and E. Vilkman, "Estimation of the glottal pulseform based on discrete all-pole modelling," in *Third International Conference on Spoken Language Processing ICSLP 94*, Yokohama, Japan, 1994, pp. 1619-1622.

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

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

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

[22] D. Ellis, "Spectral warping of LPC resonance models. https://www.ee.columbia.edu/~dpwe/resources/matlab/polewarp/ (Accessed: 12 June 2019)," 2004.

[23] A. Härmä, M. Karjalainen, L. Avioja, V. Välimäki, U. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio application," *Journal of the Audio Engineering Society,* vol. 48, pp. 1011–1031, 2000.

[24] A. Murphy, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "The role of voice quality in the perception of prominence in synthetic speech," in *Interspeech 2019*, Graz, Austria, 2019, pp. 2543-2547.

[25] R. J. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*, 2 ed. San Diego: Thomson Learning, 2000.

[26] R. D. Kent and C. Read, *The acoustic analysis of speech*, 2 ed. United Kingdom: Thomson, 2002.

[27] 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.

[28] H. M. Hanson, "Glottal characteristics of female speakers: acoustic correlates," *Journal of the Acoustical Society of America,* vol. 101, pp. 466-481, 1997.

[29] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, "Habitual use of vocal fry in young adult female speakers," *Journal of Voice,* vol. 26, pp. e111-e116, 2012.

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

[31] 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.

[32] S. Chuenwattanapranithi, Y. Xu, B. Thipakorn, and S. Maneewongvatana, "Encoding emotions in speech with the size code. A perceptual investigation," *Phonetica,* vol. 65, pp. 210-230, 2008.

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

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

[35] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software,* vol. 82, pp. 1-26, 2017.

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

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