Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Affect Expression: Global and Local Control of Voice Source Parameters

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

Phonetics and Speech Laboratory
School of Linguistic, Speech and Communication Sciences, Trinity College Dublin
murpha61@tcd.ie, yanushei@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

Abstract

This paper explores how the acoustic characteristics of the voice signal affect. It considers the proposition that the cueing of affect relies on variations in voice source parameters (including f_0) that involve both global, uniform shifts across an utterance, and local, within-utterance changes, at prosodically relevant points. To test this, a perception test was conducted with stimuli where modifications were made to voice source parameters of a synthesised baseline utterance, to target *angry* and *sad* renditions. The baseline utterance was generated with the ABAIR Irish TTS system, for one male and one female voice. The voice parameter manipulations drew on earlier production and perception experiments, and involved three stimulus series: those with global, local and a combination of global and local adjustments. 65 listeners judged the stimuli as one of the following: *angry, interested, no emotion, relaxed* and *sad*, and indicated how strongly any affect was perceived. Results broadly support the initial proposition, in that the most effective signalling of both *angry* and *sad* affect tended to involve those stimuli which combined global and local adjustments. However, results for stimuli targeting *angry* were often judged as *interested*, indicating that the negative valence is not consistently cued by the manipulations in these stimuli.

Index Terms: voice quality, speech synthesis, paralinguistics, voice source, affect, prosody

1. Introduction

The signalling of affect is known to depend to a large extent on the speaker's modulation of tone of voice, i.e. the dynamic variation of the voice (phonatory quality and f_0). The present paper examines two possible ways in which such modulations may contribute to the signalling of affect: one involves changes at a global level, across the entire utterance; another involves changes at a local level, i.e. parameter shifts that are sensitive to the prosodic structure of the utterance. This question is explored through a listening test where different parameters of the voice are manipulated in a synthesised utterance in directions that are intended to signal *angry* or *sad* emotion. This experiment aims to extend our understanding of how voice modulations communicate affect. It also aims to provide an initial, simple model of the voice transforms that can alter a neutral utterance to render it more *sad* or *angry* in synthetic speech.

A perception experiment was designed accordingly, involving both global and local changes to an utterance, synthesised using the ABAIR Irish (Gaelic) TTS voices [1]. The voice parameters manipulated included f_0, the global waveshape parameter R_d, and the excitation strength parameter E_e (for details of these parameters, see further Section 2.1.2 and [2]. The modifications were guided by previous production and perception experimentations, such as [3-6]. The number and types of adjustments were intentionally limited, as a first step towards a more elaborated model for controlling voice parameters in TTS.

There is an extensive body of research on the vocal correlates of emotion. Publications by Scherer and colleagues, e.g., [7-11] have shed considerable light on how changes in the level and dynamic range of f_0 and intensity are involved in affect signalling. These are presented in global terms, insofar as they do not provide information on whether/how such dynamic changes relate to the prosodic constituents of the utterance. Furthermore, in many of the early studies, the crucial information on voice quality was not available, and perception experiments revealed that varying f_0 and intensity without varying voice quality did not successfully cue affect.

Some of our experiments, e.g., [3, 4, 12] have specifically focused on how voice quality maps to affect, and on how voice source parameters combine with f_0 in signalling emotion, mood and attitude. Parallel research has also demonstrated that considerable voice source variation is a dimension of linguistic (non-affective) prosody, largely omitted from mainstream analysis [13]. A holistic model of voice-prosody is proposed [14, 15] that would encompass all dimensions of the voice, examining both linguistic and paralinguistic dimensions of prosody within a single analytic framework. This should yield a fuller understanding of the many dimensions of intonational meaning and of the acoustic parameters that express them.

In this paper we propose that affect-related modulations do not simply involve global shifts across the utterance but also local modulations that are sensitive to the internal prosodic structure of utterances. To test this, the present experiment explores the contribution of global and local manipulations to voice parameters in an utterance.

A finding in earlier perception experiments such as [3, 4] is that there is no one-to-one mapping between a given voice quality and affect. A particular voice quality (and f_0 contour) can potentially be associated with a number of affects, differing in activation level and valence. Thus, tense voice with dynamically varying f_0 was associated both with *angry* and *interested* – both involving high activation, but differing in valence. Similarly, a lax-creaky voice quality was associated not only with *sad* but also with low activation states such as *relaxed* and *intimate* which differ in valence. For this reason, although the stimuli in this experiment are designed to target the affects *angry* and *sad*, listeners were invited to judge them in terms of five possible responses including *angry, interested, relaxed, sad* and *no emotion*. The inclusion of *interested* and *relaxed* among the possible choices, should provide an indication as to whether the states *angry* or *sad* are unambiguously cued, as opposed to a more generalised state of high (or low) activation.

10.21437/SpeechProsody.2022-107

The use of synthetic speech as the baseline for this experiment was also motivated by practical considerations. The research group has developed Irish language TTS for male and female voices [1] which are increasingly being deployed in applications such as interactive educational games and learning platforms [16, 17]. For such applications, it is hoped to use the knowledge gleaned from these experiments to control parameters in our TTS voices in a reasonably simple way, so as to effect changes in the emotional colouring of utterances as appropriate to the context of the game, e.g., towards *angry* or *sad*. As mentioned above, the global and local manipulations of source parameters carried out in the present experiment were intentionally constrained in a way that would facilitate manipulation in our TTS voices, serving as a potential model for controlling the speech output. The experiment uses as a starting point an Irish sentence, synthesised using the ABAIR TTS system for a male and female voice.

2. Methods

As a way of exploring the proposition that a combination of global and local changes in voice parameters are involved in affect signalling, the present perception experiment compares listener judgements of affect for three stimulus types. These were (i) *global* stimuli, where voice parameters were altered utterance-wide, (ii) *local* stimuli, where voice parameter changes were confined to the nuclear contour, and (iii) combined *global+local* stimuli, where both these global and local changes to the voice parameters were included. The hypotheses tested are that (a) the affects *angry* and *sad* would be cued by these stimuli; (b) the combined *global+local* stimuli would be the most effective in cueing these affects, and (c) the cueing of *angry/sad* would be unambiguous (i.e., responses would be considerably higher for these than for *interested/relaxed*).

2.1. Synthetic stimuli

2.1.1. Sentence and TTS voices used for baseline stimuli

Two synthesised utterances were the starting point for stimulus construction. These were of a sentence generated using two of the ABAIR DNN synthetic voices [18], one male and one female, both of the Kerry (Dingle) dialect of Irish:

Beidh an lá go hálainn amárach.

'Will be the day lovely tomorrow.' (word gloss)

'The day will be lovely tomorrow.' (translation)

2.1.2. Voice parameters controlled

The voice source characteristics of the baseline synthetic stimuli were initially analysed and subsequently modified to target either the affect *sad* or *angry*, using global, local or combined global+local manipulations. As mentioned, these were guided by prior analysis of production data as well as earlier perception experiments [3-6]. The voice source parameters modified included f_0, R_d and E_e (see Fig. 1).

R_d is a global waveshape parameter which has been shown to capture voice source variation in the tense-lax dimension. Values typically range between 0.5 (tense voice) to 2.5 (breathy voice). R_d is derived from f_0, E_e and U_p as follows: $(1/0.11)\times(f_0 \cdot U_p/E_e)$.

E_e is a measure of the strength of the glottal excitation, measured as the negative amplitude of the differentiated glottal flow at the time point of maximum waveform discontinuity.

U_p is the peak amplitude of the glottal flow pulse. It is not directly manipulated here, but varies depending on the settings of the other three parameters, $f_0 \cdot R_d$ and E_e.

Note that U_p/E_e is equivalent to the glottal pulse declination time T_d during the closing phase of the glottal cycle. The scale factor (0.11^{-1}) makes the numerical value of R_d equal to the pulse declination time in milliseconds when f_0 is 110 Hz [2].

By changing R_d, other parameters of the glottal pulse such as R_a and R_k also vary, and these changes can be predicted from R_d. To synthesise the LF model glottal waveform, data for the full set of LF model parameters are required and were calculated from R_d using the parameter correlations presented in [2].

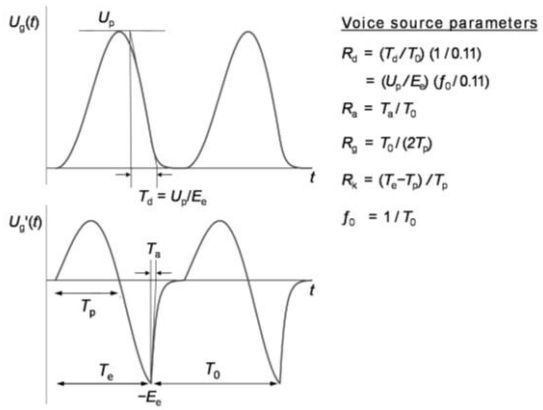

Figure 1: *The LF model and voice source parameters (adapted from [2]).*

2.1.3. Voice analysis and synthesis

The synthetised sentences for the baseline were analysed using the GlórCáil system [19] which enables analysis and resynthesis of speech and control of voice source parameters using an interactive GUI. Vocal tract and voice source parameters were estimated for these sentences, using automatic inverse filtering GFM-IAIF [20, 21], and source parameterisation using [22]. For resynthesis, concatenated LF model [23] pulses are used in place of the original voice source, which are then filtered by an approximation of the vocal tract filter. Aspiration noise is also added to the synthetic source signal using the method described in [24] where the amplitude of the aspiration noise varies as a function of the glottal pulse shape. Modelling the source in this way allows the user 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. For more detail on the system see [25].

2.1.4. Manipulations for generating affect-targeted stimuli

Parameters were modified to create affect-targeted stimuli based on prior analyses and perception experiments. The scale factors for the manipulations were guided by voice source data obtained in an earlier study for a male speaker who produced an utterance with different affective renditions, including *sad*, *angry* as well as a neutral version [6]. Table 1 summarises the kinds of manipulations carried out.

The baseline stimuli: the sentences generated for the male and female TTS voices were analysed and the voice parameters (R_d, E_e and f_0) were scaled to the mean and standard deviation of the neutral utterance in [6].

Table 1: *Parameter manipulations for affect colouring.*

Target affect	Global (entire utterance)	Local (nuclear contour)	Combined: global + local
Neutral baseline	scaled according to production data		
Angry voice	neutral + f_0 raised + R_d lowered	neutral + higher f_0 peak + rapid f_0 fall + corresponding changes in R_d and E_e	global-*angry* + local-*angry*
Sad Voice	neutral + f_0 lowered + R_d raised	neutral + flatter f_0 peak + flatter E_e + corresponding changes in R_d	global-*sad* + local-*sad*

The global stimuli involved further changes to R_d, E_e and f_0, applied to the baseline stimuli, altered to match the means and standard deviations of either the *sad* or *angry* production data in [6]. These settings are provided in Table 2.

The local stimuli: involved changes to the nuclear contour (the nuclear syllable and the post-nuclear unaccented syllable). The local manipulation for *angry* involved raising the f_0 peak of the nuclear syllable by 20% followed by a downwards ramp to 40% of the original f_0 peak value across the remainder of the utterance (to the end of the postnuclear syllable). This was accompanied by a lowering of R_d (towards a tenser voice quality) by 20% on the nuclear syllable, followed by a ramping to 60% of the original R_d value across the remainder of the utterance. The local manipulation for *sad* involved flattening both f_0 and E_e across the nuclear syllable and the remainder of the utterance, along with a flattening and raising of R_d, resulting in a more lax voice quality.

The combined global+local stimuli involved altering the baseline to include both global and local changes towards *angry* and *sad* target affects.

Table 2: *Global parameter manipulations.*

Affect	Gender	Mean f_0 (Hz)	SD f_0 (Hz)	Mean R_d	SD R_d	Mean E_e (dB)	SD E_e (dB)
Neutral	Male	96	9	1.05	0.21	69	3
	Female	169	29	"	"	"	"
Angry	Male	127	18	0.82	0.20	72	4
	Female	224	58	"	"	"	"
Sad	Male	94	2	1.81	0.22	62	6
	Female	162	6	"	"	"	"

There were 14 stimuli overall: 2 voices (male, female) × 2 target affects (sad, angry) × 3 types of manipulation (global, local, combined global+local) + 2 neutral baselines (one for each voice, globally scaled according to production data).

2.2. Listening test

The stimuli were presented to the participants, who were all Irish speakers (n = 65), in an online listening test. The listeners first listened to all the stimuli to get some idea of the range of variation. They were then instructed to listen to each stimulus once and decide how the speaker sounds (multiple choice: *sad, angry, interested, relaxed,* or *no emotion*). If an emotion was deemed present, they were asked to indicate the strength of the emotion (from weak to strong, on a 5-point scale). The instructions were given in Irish. The 14 stimuli were presented to the participants five times in random order. Four additional stimuli were added to the beginning of the test to serve as a practice run: these were discarded from the analysis.

3. Results and discussion

Fig. 2 shows the frequency of affect judgements (in %) for the female and male stimuli, with absolute values on the left and values relative to the baseline on the right. Results for the baseline stimuli are shown with a dashed grey line in the left panel. We used Pearson's Chi-squared test for count data. To compare individual stimuli to the baseline, standardised residuals were used for post-hoc analysis using the *chisq.posthoc.test* package [26] in *R 4.0.5* [27]. Asterisks denote responses for *angry* and *sad* that are significantly higher than the baseline (BL). The strength of the emotion is shown by the size of the data points.

(a) How successfully are *angry* and *sad* signalled?
Broadly speaking, the goal of shifting the perceived affect of an utterance towards *angry* or *sad* appears to have been successful. Of the two targeted affects, *sad* emerges more clearly for both the female and male stimuli. In the case of *angry*, judgements are divided between *angry* and *interested*. In absolute terms (left panel), the *angry* responses to the male stimuli were the most frequent, marginally higher than *interested*. For the female stimuli *angry* responses are less frequent than *interested*.

However, the absolute values are skewed insofar as the baseline stimuli (dashed lines) are not affect-neutral. As the principal focus here is to ascertain the extent to which the manipulations for the three stimulus types can alter the perceived affect of an utterance, the right panel shows values normalised to the baseline. Note that from this perspective, the pattern of *angry* responses emerges more strongly and are very similar for the male and female voices. The single highest affective score (regardless of stimulus type) indicates a clear shift from the baseline towards *angry* for both male and female stimuli.

Among the male stimuli, when seen in terms of the shifts from the baseline, *angry* is now clearly the most frequent response (for the combined stimulus) showing a 44% shift towards *angry* relative to the baseline (+44% re BL) [χ^2 = 270.72, df = 4, p < .001]. The highest value for *interested* responses (for the global stimulus) was +27% re BL.

For the female stimuli, although angry responses are now also the most frequent (the combined stimulus, +30% re BL and significantly different from it [χ^2 = 165.93, df = 4, p < .001], this is only marginally more frequent than the *interested* responses for the global stimulus (+26% re BL).

(b) Are the combined stimuli the most effective?
The initial hypothesis was that the stimuli combining both global and local modulations would be the most effective in cueing the *sad* and *angry* affects. Results broadly support this.

For the angry-targeted stimuli, male and female, the combined stimulus (orange data points) produced the highest frequency of *angry* responses. The male combined stimulus yielded significantly more frequent responses than either the global or the local stimuli (by about 23% in both cases). For the female voice, the combined stimulus also yielded significantly more frequent *angry* responses than the other two (17% increase relative to the global and/or local, χ^2 = 24.42, df = 4, p < .001). For both male and female stimuli, there appears to be an additive effect by combining global and local voice source manipulations.

Results for the *sad*-targeted stimuli were less clear-cut. In the case of the male stimuli, the combined stimulus yielded the

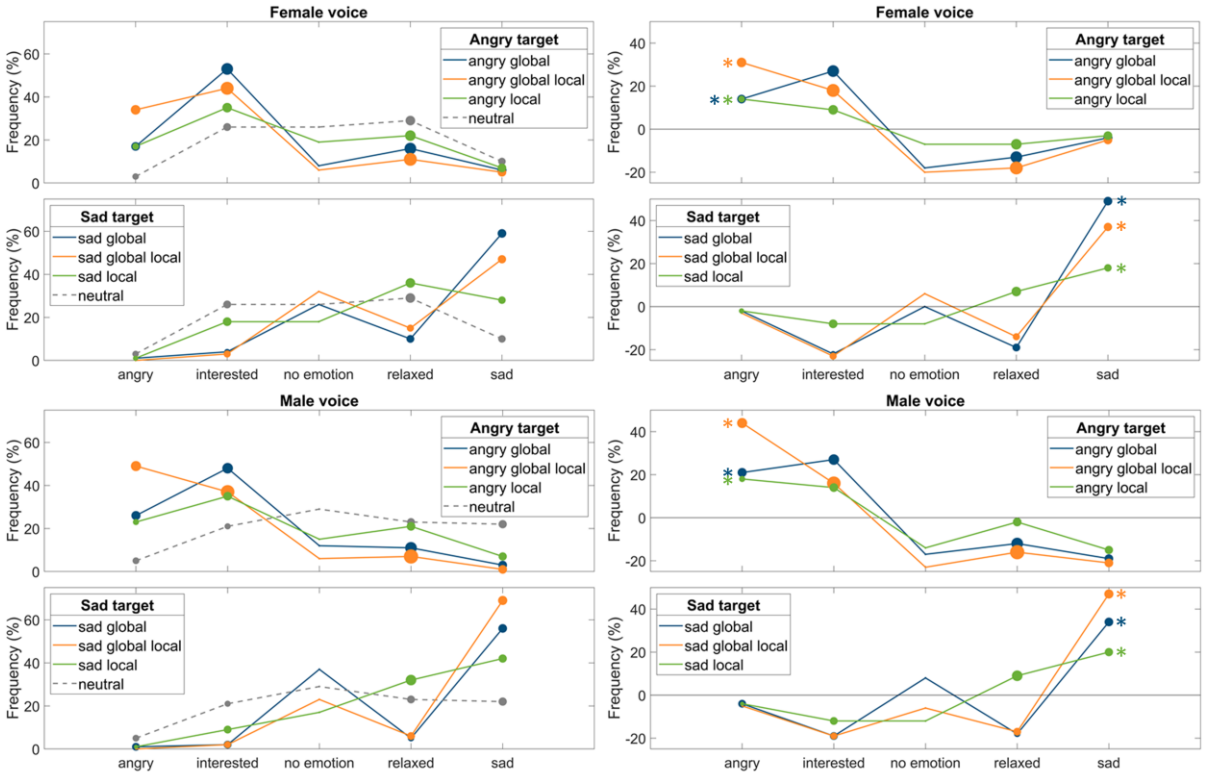

Figure 2: *Frequency of affects perceived (%) as absolute values (left) and normalised to the baseline stimulus (right). Asterisks indicate significant difference from the baseline.*

most frequent *sad* responses (+47% re BL) [χ^2 = 182.35, df = 4, p < .001] a value significantly higher than for the local (+20% re BL) or global (+34% re BL) stimuli.

The responses for the female stimuli targeting *sad* did not follow suit. Here, the global manipulation (+48% re BL) yielded the most frequent *sad* responses, although this is not significantly different to the response for the combined stimulus (+37% re BL) [χ^2 =14.18, df = 4, p = .007; posthoc test of residuals p = .06]. The local stimulus shows a relatively weaker but significant effect (+17% re BL). In this case, combining the local and global manipulations reduces rather than enhances the *sad* response.

(c) Is the signalling of angry/sad unambiguous?

Results here are mixed. For the stimuli targeting *sad*, the *relaxed* option is rarely chosen, suggesting a rather unambiguous cueing. However, for stimuli targeting *angry*, there was considerable ambiguity, especially in the case of the female stimuli. For the male voice, when one considers the highest affect-shifting stimulus relative to the baseline – the combined stimulus – responses for *angry* (+44% re BL) are higher than for *interested* (+16% re BL) but it is clear that *interested* responses are relatively frequent (+27% re BL for the global stimulus). This ambivalence is even more pronounced for the female stimuli. We conclude that *angry* is not unambiguously cued across these stimuli.

The strength of the affect (shown by the size of the data points) differs little, being mostly clustered around 3 (out of 5), but *interested*, when chosen, is more strongly rated than *angry*. Listeners were thus clearly divided between an interpretation of these stimuli as exhibiting either a moderate degree of anger or a somewhat higher degree of interest.

4. Conclusions

These results support the initial hypothesis that affect signalling entails a combination of global, utterance-wide shifts in voice parameters, and local changes that are sensitive to prosodic structure. As such, it invites a broader perspective on the nature of prosody, where we would argue, a holistic approach is needed to understand the forms and functions of prosody in speech communication – an approach that would include all the acoustic dimensions of the voice (voice quality and f_0) in a way that allows examination of how they shape the many dimensions of prosodic meaning [13-15].

This experiment is intended to provide a preliminary model to control voice parameters in our TTS system, to generate utterances with some changes in perceived affect. We hope to proceed with more complex models that include further dimensions of voice quality, such as creaky and whispery voice, using them to explore how local prosodic structure may shape the overall vocal changes that signal affect. Temporal aspects, and particularly the alignment of f_0 and source parameters, will also require investigation. A practical application such as an interactive game may open up further avenues of research, including exploration of how both the inherent semantic content of an utterance and the pragmatic context of the game impact on the affective interpretation of vocal cues.

5. Acknowledgements

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 *Straitéis 20 Bliain don Ghaeilge*.

6. References

[1] 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. 2021, pp. 2113-2117.

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

[3] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Comm.*, vol. 40, no. 1-2, pp. 189-212, 2003.

[4] Yanushevskaya, C. Gobl, and A. Ní Chasaide, "Cross-language differences in how voice quality and f_0 contours map to affect," *J. Acoust. Soc. Am.*, vol. 144, no. 5, pp. 2730-2750, 2018.

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

[6] Yanushevskaya, C. Gobl, and A. Ní Chasaide, "Voice parameter dynamics in portrayed emotions," in *6th International Workshop on Models and Analysis of Vocal Emissions for Biometrical Applications (MAVEBA 2009)*, Florence, Italy, 2009, pp. 21-24.

[7] K. R. Scherer, D. R. Ladd, and K. E. A. Silverman, "Vocal cues to speaker affect: testing two models," *J. Acoust. Soc. Am.*, vol. 76, no. 5, pp. 1346-1356, 1984.

[8] D. R. Ladd, K. E. A. Silverman, F. Tolkmitt, G. Bergmann, and K. R. Scherer, "Evidence for the independent function of intonation contour type, voice quality, and f_0 range in signaling speaker affect," *J. Acoust. Soc. Am.*, vol. 78, no. 2, pp. 435-444, 1985.

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

[10] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Pers. Soc. Psychol.*, vol. 70, no. 3, pp. 614-636, 1996.

[11] K. R. Scherer, "Comment: Advances in studying the vocal expression of emotion: Current contributions and further options," *Emot. Rev.*, vol. 13, no. 1, pp. 57-59, 2021.

[12] 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.

[13] A. Ní Chasaide, I. Yanushevskaya, J. Kane, and C. Gobl, "The Voice Prominence Hypothesis: the interplay of f_0 and voice source features in accentuation," in *INTERSPEECH 2013*, Lyon, France, 2013, pp. 3527-3531.

[14] A. Ní Chasaide and C. Gobl, "Voice quality and f_0 in prosody: towards a holistic account," in *Speech Prosody 2004*, Nara, Japan, 2004, pp. 189-196.

[15] A. Ní Chasaide and C. Gobl, "Decomposing linguistic and affective components of phonatory quality," in *INTERSPEECH 2004*, Jeju Island, Korea, 2004, pp. 901-904.

[16] A. Ní Chasaide *et al.*, "Leveraging phonetic and speech research for Irish language revitalisation and maintenance," in *The XIX International Congress of Phonetic Sciences*, Melbourne, Australia, 2019, pp. 994-998.

[17] N. Ní Chiaráin and A. Ní Chasaide, "Evaluating text-to-speech synthesis for CALL platforms," in *Antwerp CALL 2014: International CALL Research Conference*, Antwerp, 2014: University of Antwerp, pp. 104-110.

[18] ABAIR.ie – The Synthesiser for Irish, https://abair.ie

[19] A. Murphy, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Testing the GlórCáil system in a speaker and affect voice transformation task," in *Speech Prosody 2020*, Tokyo, Japan, 2020, pp. 950-954.

[20] O. Perrotin and I. V. McLoughlin, "On the use of a spectral glottal model for the source-filter separation of speech," *CoRR*, vol. abs/1712.08034, 2017.

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

[22] J. Kane, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Exploiting time and frequency domain measures for precise voice source parameterisation," in *Speech Prosody 2012*, Shanghai, China, 2012, pp. 143-146.

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

[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] A. Murphy, *Controlling the Voice Quality Dimension of Prosody in Synthetic Speech using an Acoustic Glottal Model*, unpublished PhD Thesis, Trinity College Dublin, 2020.

[26] D. Ebbert, "chisq.posthoc.test: a post hoc analysis for Pearson's chi-squared test for count data," *R package version 0.1.2*, 2019.

[27] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria, 2021.