Integrating a Voice Analysis-Synthesis System with a TTS Framework for Controlling Affect and Speaker Identity

Andy Murphy  
Phonetics and Speech Laboratory  
Trinity College Dublin  
Dublin, Ireland  
murphyah@tcd.ie

Irena Yanushevskaia  
Phonetics and Speech Laboratory  
Trinity College Dublin  
Dublin, Ireland  
yanusheiv@tcd.ie

Ailbhe Ní Chasaide  
Phonetics and Speech Laboratory  
Trinity College Dublin  
Dublin, Ireland  
anichsid@tcd.ie

Christer Gobli  
Phonetics and Speech Laboratory  
Trinity College Dublin  
Dublin, Ireland  
cegobli@tcd.ie

Abstract—This paper reports an experiment exploring how a voice analysis-synthesis system, GlórCáil, can be used to add expressiveness to the synthetic voice in TTS systems. This implementation focuses on the Irish ABAIR text-to-speech (TTS) voices, where such voice control would facilitate many current/envisaged applications. GlórCáil allows voice control of synthesized speech and for this experiment was integrated into a DNN-based TTS framework. Utterances were generated with $f_0$, voice quality and vocal tract parameter manipulations targeting shifts in speaker identity and in the affective coloring of utterances. Scaling factors used for the manipulations were suggested in an earlier study. They involved global changes without sentence-internal dynamic variation, with a view to ascertain whether such global shifts might alter listeners’ perception of speaker identity and affect. Results demonstrate affect shifts compatible with expectations. However, there were confounding factors. The female-child voices were poorly differentiated, which was expected given the similarity in the scaling factors used. The affect transformations suggest the baseline voice used had an intrinsically sad quality so that there is weak differentiation between the sad and no emotion stimuli. Male angry voice was the least successful suggesting that dynamic, within-utterance variation is essential for the signaling of certain affects.

Keywords—speech synthesis, voice quality, voice transformation, affect, speaker characteristics

I. INTRODUCTION

Building on past research on the voice in human and synthetic speech, this paper explores an approach to controlling synthesis parameters in TTS, in order to alter the perceived gender/age and affect. Gaining control of the voice source is a vital step towards attaining this goal. One approach to this problem is to implement an acoustic model that closely matches the natural glottal source [1], [2]. However, such models involve complex parametric schematics making them difficult to control by non-expert users.

The GlórCáil system for voice analysis-synthesis [3] was developed to allow for the transformation of voice quality in synthetic speech using a small set of control parameters through an interactive GUI that can be easily and intuitively operated by non-expert users. This system incorporates knowledge of the speech production process and, in particular, uses an acoustic glottal model-based excitation that closely models the natural glottal source. At the analysis stage, the system extracts estimates of the vocal tract and voice source parameters. Speech is then resynthesized using an acoustic glottal source model in place of the original glottal source signal. The interactive GUI allows the user to manipulate vocal tract scaling factors and voice source parameter contours in an utterance, listen to the results of manipulations and make further changes as desired to achieve variations to the perceived voice quality. The system is explained in some detail in Section II.B below.

The system’s ability to alter such speaker characteristics as gender, age and affective coloring of speech was explored in a voice transformation task [3]. The task required participants to manipulate scaling factors linked to the system’s control parameters, reflecting changes in voice quality, $f_0$ and vocal tract length, in order to alter the perceived speaker identity and affect. The results of the study suggested that the system’s control parameters were useful to this end.

Such voice control is a matter of particular interest in relation to text-to-speech (TTS) systems which have been developed for the main Irish dialects [www.abair.ie]. These voices are increasingly being deployed in applications, such as interactive multimodal educational games [4], where it is highly desirable to have the facility to alter the age/gender of the voice, and especially, the expressive coloring of utterances.

In this paper, the GlórCáil system is integrated into a deep neural network (DNN) TTS framework (GlórCáil-DNN), allowing manipulation of the above parameters in the speech output. An experiment is reported here, where stimuli are generated utilizing the scaling factors obtained in [3], as an initial trial of how the ABAIR DNN voices might be transformed.

The manipulations carried out in the present experiment are of a global kind, where uniform scaling of glottal source parameters and a vocal tract warping function are applied to an entire utterance. Although earlier work demonstrates that prosody-sensitive, sentence-internal changes in voice dynamics are likely to be important in the signaling of certain affects [5], the aim of this experiment is to explore, as a first step, to what extent these simple shifts can alter the speaker identity and affective colouring. It has been shown that simple global shifts in source and filter parameters can effect convincing male-to-female voice transformation [6]. Furthermore, the multidimensional wave shape parameter $R_d$ [7] has been shown to be able to capture voice quality variation on the tense-lax dimension and consequently important in the signaling of affect [8].

II. METHODS AND MATERIALS

The following sections describe how the stimuli for the experiment were generated by a baseline statistical parametric
speech synthesis system and the GlórCáil-DNN system. Details concerning the listening test are also provided.

A. Baseline system

The baseline system used in this work was built using the nmmkwiiv Python library [9]. This library makes prototyping of speech synthesis systems fast and easy due to its simplicity and transparent design. Some of the system’s scripts are based on Merlin [10] speech synthesis system demo scripts. The WORLD vocoder [11] is used to extract log \( f_0 \), spectral and aperiodicity parameters and to resynthesise speech from parameters generated by the DNN component.

B. The GlórCáil system

The GlórCáil system is an application developed for the analysis and resynthesis of speech with a particular focus on control of voice source parameters using an interactive GUI. It is implemented in the MATLAB environment [12]. Vocal tract and voice source parameters are first estimated during the analysis stage using Glottal Flow Model-based Iterative Adaptive Inverse Filtering (GFM-IAIF) [13], [14] and source parameterisation using [15]. Speech is then resynthesized using concatenated Liljencrants-Fant (LF) model [16] pulses (in place of the original glottal 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 [17] where the magnitude of the aspiration noise varies as a function of the glottal pulse shape. Modeling the source in this way allows the user to manipulate voice source parameter contours in an interactive GUI, listen to the results and make further changes to the voice quality of the utterance.

1) GlórCáil analysis stage

A flowchart of the analysis stage is shown in Fig. 1. The analysis involves an initial polarity check, followed by \( f_0 \) estimation, detection of glottal closure instant (GCI) locations and voiced/unvoiced (VUV) regions of a speech signal. Inverse filtering is performed on the voiced regions to obtain an estimate of the glottal source. This signal is parameterized by fitting the LF-model to each differentiated glottal flow (DGF) pulse. LPC analysis is performed on unvoiced regions of speech to obtain an estimate of the vocal tract transfer function. More detailed descriptions of the different components follow below.

![Flowchart of the GlórCáil analysis stage.](Image)

The first step of the GlórCáil analysis stage is to check the polarity of the signal being analysed. The use of varying microphones and recording equipment results in speech data with different polarities. Correcting the polarity of a signal is important due to the fact that the accuracy and performance of many analysis techniques may be negatively affected if the polarity of the signal is inverted. This is an important step as the method for modelling the DGF [18] uses a polarity-sensitive time domain measure. The polarity check is carried out using the method described in [19]. It determines the polarity of a signal by first calculating the residual of the signal and a rough estimation of the glottal flow derivative. The skewness of the sample distributions of each signal is then calculated and subtracted from each other. The sign of the result is the polarity of the signal, with a negative value signifying an inverted or negative polarity, meaning that the signal must be multiplied by -1 to correct it.

The next step is to estimate \( f_0 \), GCI locations, and VUV regions. This is carried out using the REAPER programme [20].

Inverse filtering is then carried out on a frame-by-frame basis using the modified version of IAIIF, GFM-IAIF with a frame length of 25 ms, a frameshift of 5 ms, and an order determined by \( F_s/1000 + 4 \), where \( F_s \) is the sampling frequency in hertz. Discrete all-pole (DAP) modelling is used in place of LPC modelling in the GFM-IAIF process as it provides better estimations of the vocal tract transfer function [21]. This analysis provides estimations of the DGF 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 a statistical parametric speech synthesis system.

LSFs are the roots of the line spectral pair (LSP) polynomial described in [22] and [23]. The inverse filter resulting from an LPC analysis can be expressed in the Z domain by the \( m \)th order polynomial:

\[
A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_m z^{-m}
\]  

The LPC coefficients \((a_1, a_2, a_3, \ldots, a_m)\) are very sensitive to quantisation errors, which can lead to filter instabilities. The \( A(z) \) polynomial can instead be expressed as:

\[
A(z) = 1/2[P(z) + Q(z)]
\]

where \( P(z) \) and \( Q(z) \) are symmetric and anti-symmetric LSP polynomials respectively. These polynomials are defined as:

\[
P(z) = A(z) \left( 1 + z^{-(m+1)} \frac{A(z^{-1})}{A(z)} \right)
\]

\[
Q(z) = A(z) \left( 1 - z^{-(m+1)} \frac{A(z^{-1})}{A(z)} \right)
\]

The polynomials have the following properties when \( A(z) \) has all roots within the unit circle:

1. Their zeros all fall on the unit circle;
2. Their zeros are interlaced with each other;
3. The minimum phase property of \( A(z) \) is maintained after the zeros of \( P(z) \) and \( Q(z) \) are quantized.

The first two properties are valuable when calculating the zeros of \( P(z) \) and \( Q(z) \), and the third property guarantees the stability of the filter derived from the polynomials [23]. Once the differentiated glottal waveform has been obtained, model fitting is performed to parameterize each pulse using a method similar to that described in [18]. Each analysis frame consists
of a GCI(n) centred glottal pulse that extends from GCI(n−1) to GCI(n+1) and is windowed using a Hann window.

A range of pulses are generated using values of the global wave scheme parameter, \( R_0 \) [7], between 0.3 and 2.7 in 0.05 increments and the maximum negative value of the DGF pulse, \( E_c \), \( E_a \), as well as \( f_0 \) are used to calculate a set of default parameters to generate the corresponding LF pulse for each incremental value of \( R_0 \) within the range. The time domain and spectral (amplitude and phase) correlation coefficients of the DGF pulses, when compared with the generated LF pulses, are calculated and subtracted from 1 to give an error value for each of these measures. These error values are then sorted according to increasing error. The top ten \( R_0 \) candidates that minimise the error values are selected for the next stage of the process. A dynamic programming method [24] is used to calculate the optimum path of \( R_0 \) values through the utterance. The optimal \( R_0 \) values are then combined with their corresponding \( E_c \) and \( f_0 \) values as the source parameters. The extracted voice source parameters are then smoothed using a 5th order median filter followed by a 5th order moving average filter before being saved.

For unvoiced frames, LPC analysis is carried out to approximate the unvoiced spectrum. The gain of the unvoiced frames is measured by taking the root-mean-square of the prediction error. The unvoiced LPCs are also converted to LSFs.

2) GlórCáil synthesis stage

Once parameters have been extracted from an utterance in the analysis stage, the GlórCáil synthesis stage (see Error! Reference source not found.) can be used to resynthesise it. This stage includes an option to manipulate the voice source parameter contours of an utterance so that it can be resynthesised, and the results of the manipulations can be listened to straight away. Users can also compare a manipulated utterance with the original unmanipulated version directly in the interface.

The process of resynthesis begins by generating the voiced and unvoiced excitations. Voiced regions are defined by frames with \( f_0 \) values above and below a minimum and maximum frequency (default 50 Hz and 500 Hz respectively). The voiced excitation consists of LF pulses 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 of \( f_0 = 1/T_0 \). The period is used to define the start and end points of the pulse frame within the voiced excitation. The parameters \( E_c \), \( R_0 \) and \( f_0 \) are used to generate the full set of LF-model parameters using the correlations outlined in [7], which are used to generate the corresponding LF pulse. Amplitude modulated noise is added to the pulse according to the method described in [17]. The level of this noise is modulated according to the shape of the pulse. A pulse corresponding to a tense voice quality will have little to no added noise, while a pulse corresponding to a lax voice quality will have a considerable amount of aspiration noise added to it.

This first pulse is placed within the start and end indices that have already been defined, and 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 values of zero, so no windowing is required to prevent discontinuities in the signal.

Frames with values of \( f_0 \) below the minimum \( f_0 \) threshold are treated as unvoiced regions. White Gaussian noise is used as the excitation signal in these areas.

The excitation signals are then filtered by a filter that represents the vocal tract transfer function. The voiced excitation is filtered using the LPC filter coefficients that are converted back from the LSFs that were output from the synthesis stage. A scaling factor may 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. This approach involves substituting the original filter coefficients with first order all-pass elements using the mapping shown in (5) [25],

\[
\tilde{z}^{-1} \rightarrow z^{-1} = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}
\]

where \( z^{-1} \) is the original filter coefficient, \( \tilde{z}^{-1} \) is the transformed coefficient, and \( \lambda \) 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 method is based on the implementation provided by [26]. This ability to scale the resonant peaks 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 LPC analysis. The final speech signal is then created by overlap-adding each of the filtered frames.

C. Training

The GlórCáil-DNN and baseline systems were trained using a corpus of 1452 utterances spoken by a male speaker of Irish (Kerry dialect) in his mid-20s. The audio was recorded in a semi-anechoic chamber using a B&K microphone at 44.1 kHz with a bit depth of 16, then resampled to 16 kHz prior to training. Two Bidirectional Long Short-Term Memory (BLSTM) networks [27] were used, one to model the duration and one the acoustic features. Each network consisted of three hidden layers with 512 neurons in each layer.

D. Stimuli for the listening test

[Sentences] Two sentences that were not used in the training of the systems were used as the text input to generate the stimuli for this experiment:

![Flowchart of the GlórCáil synthesis stage.](Image)
Sentence 1:
Tá cír sa nu a sheoladh i gColáiste na Tríonóide, Baile Átha Cliath.
[IPA: tə ɾˠauŋɡˠə bˠəɾˠə əɾˠə nˠoːdˠə aːhə clɪə]
‘Is course new being launched in College of Trinity, Dublin.’
(word gloss)
‘A new course is being launched in Trinity College Dublin.’
(translation)

Sentence 2:
Tá cíš na nua á sheoladh i gColáiste na Tríonóide, Baile Átha Cliath.
[IPA: tə kʰʊɾˠə əɾˠə nˠoː dˠə əɾˠə nˠoːdˠə aːhə clɪə]
‘Is class new being launched in College of Trinity, Dublin.’
(word gloss)
‘A new course is being launched in Trinity College Dublin.’
(translation)

[Scaling factors] The stimuli were synthesised using the scaling factors obtained in a preliminary study with eight participants following the methodology in [3]. The participants manipulated scaling factors linked to GlórCáil control parameters reflecting changes in voice quality, \( f_0 \) and vocal tract length in order to make a synthetic utterance sound like a target speaker (man, woman, child) with an affective coloring (sad, angry, no emotion). These scaling factors are shown in Table I.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Affect</th>
<th>( R_1 )</th>
<th>( f_0 )</th>
<th>( VT )</th>
</tr>
</thead>
<tbody>
<tr>
<td>child</td>
<td>angry</td>
<td>0.66</td>
<td>2.28</td>
<td>-0.057</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>1.65</td>
<td>2.11</td>
<td>-0.062</td>
</tr>
<tr>
<td></td>
<td>no emotion</td>
<td>1.09</td>
<td>1.97</td>
<td>-0.060</td>
</tr>
<tr>
<td>woman</td>
<td>angry</td>
<td>0.77</td>
<td>1.98</td>
<td>-0.050</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>1.63</td>
<td>2.01</td>
<td>-0.014</td>
</tr>
<tr>
<td></td>
<td>no emotion</td>
<td>1.13</td>
<td>1.59</td>
<td>-0.047</td>
</tr>
<tr>
<td>man</td>
<td>angry</td>
<td>0.63</td>
<td>0.98</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>1.56</td>
<td>0.94</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>no emotion</td>
<td>1.02</td>
<td>0.98</td>
<td>0.012</td>
</tr>
</tbody>
</table>

[DNN synthesis] The scaling factors of \( R_1 \) and \( f_0 \) as well as an overall vocal tract warping coefficient (VT) shown in Table I. were used in the GlórCáil-DNN system to generate 18 stimuli based on sentences 1 and 2 (nine versions of each sentence). The original voice source parameter contours and vocal tract warping factor of sentences 1 and 2 were scaled in such a way as to correspond to each of the nine combinations of target speaker (child, woman and man) and affective states (angry, sad and no emotion). Two additional versions of sentences 1 and 2 were synthesized using the baseline system (Section 2.1), making the total number of stimuli 20.

The scaling factors used in the generation of stimuli of different affect by gender combinations were obtained in the manipulation task using a different baseline voice [3]. By applying them directly to transform the voice of a different baseline speaker we are making assumptions that the same or similar result will be found. Although both speakers were Irish males of similar age and similar speaking \( f_0 \), there were differences in voice quality.

Preliminary auditory analysis was conducted by the authors to confirm that the synthesized utterances were adequate in terms of speaker and affect differentiation and naturalness.

E. Listening test
The listening test was carried out using online survey software [28]. Forty synthetic stimuli (two repetitions of the set of 20 stimuli) were presented in random order to 20 participants. The participants were informed that they would hear a number of different utterances, and that they could listen to each audio file as many times as they wished in order to answer the following questions for each stimulus:

4. Who is the speaker? [participants chose from radio buttons with the options child, woman or man]
5. How does the speaker sound? [participants chose from radio buttons with the affective labels angry, sad and no emotion];
6. To what extent? [a five-point scale ranging from ‘Not at all’ to ‘A lot’];
7. How natural is the speech? [a five-point scale ranging from ‘Not at all natural’ to ‘Very natural’].

Based on the similarity of the scaling factors between woman and child voices (TABLE I.) we could predict that the participants would be less likely to distinguish between the woman and child voices. Furthermore, the identification of stimuli as angry would be more difficult than sad and no emotion. Although non-optimal in these ways, it was felt that nonetheless these scaling factors would suffice for a first proof-of-concept demonstration of the system’s use.

III. RESULTS

A. Speaker and affect identification
Results of the speaker identification are shown as a confusion matrix in Fig. 3 (left panel). As expected, the man stimuli were always correctly identified as spoken by a male. The child stimuli were identified as such in 60.5% of the cases, and the woman stimuli were the least frequently identified as the intended target (in 42.1% of the cases). Both woman and child stimuli were frequently confused, e.g., the woman stimuli were identified as child in over 50% of the cases. This was not unexpected based on the preliminary auditory analysis of the stimuli by the authors. Although the transformation scaling values used to generate the stimuli in this study were obtained from the interactive manipulation task in a listening test (described in [3] and in section II.D), the scaling factors for woman and child ones were often very similar.

Fig. 3 (right panel) also shows the results of affect identification for the selected cases where the target speaker was identified correctly. The sad stimuli were most frequently identified as intended, in 64.5% of the cases. The no emotion stimuli were correctly identified in 45% of the cases and were mainly confused with sad (in 46.4% of the cases). The identification of the angry stimuli was not consistent: those were more readily identified as no emotion (37.2%) and were almost equally likely to be identified as either angry (32.1%) or sad (30.8%).

B. Affect identification per speaker
The affect identification results on a per-speaker basis are shown in Fig. 4.
[Child stimuli] Target affect identification was above chance but never reached 50% accuracy. No emotion child was identified as such in 47.6% of the cases, but was frequently confused with sad child (38.8% cases). Sad child was identified as no emotion in 48% of the cases. Angry child was the least frequently identified as intended (in 43.6% of the cases).

[Woman stimuli] Sad woman was identified as intended in 60.5% of the cases. However, there seems to be a poor differentiation of sad and no emotion in woman as the latter was often perceived as sad (60.6% of the cases) and less frequently as no emotion (30.3% of the cases). The identification of angry woman voice was similar to that of angry child: the voice was identified as angry in 44% of the cases, as no emotion in 36% and as sad in 20% of the cases.

[Man stimuli] No emotion and sad man were perceived as intended in higher percentage of the cases compared to the child and woman voices. Sad man was perceived as intended in 76.3% of the cases (suggesting optimal scaling of parameters and more natural sounding voice). No emotion was identified ‘correctly’ in 50% of the cases. The angry man voice was perceived as angry only in 19.7% of the cases (far less frequently than woman or child voice), and more readily perceived as no emotion (40.8%) or sad (39.5%) voice.

The results of the five-point naturalness rating for the responses where both speaker and affect were identified correctly are shown in TABLE III. The baseline stimuli had the highest naturalness rating at 3.96, closely followed by angry man at 3.93. Other man stimuli also had higher naturalness ratings than the woman and child stimuli: 3.81 and 3.74 for sad and no emotion respectively. Both child and woman voices had lower naturalness ratings overall, with angry stimuli having the lowest ratings: 2.75 and 3 for child and woman respectively.

### TABLE II. MEAN MAGNITUDE OF PERCEIVED AFFECT.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Emotion</th>
<th>Magnitude</th>
<th>N</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>child</td>
<td>angry</td>
<td>3.13</td>
<td>24</td>
<td>.900</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>2.64</td>
<td>19</td>
<td>1.167</td>
</tr>
<tr>
<td>woman</td>
<td>angry</td>
<td>2.36</td>
<td>11</td>
<td>.924</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>3.00</td>
<td>23</td>
<td>.905</td>
</tr>
<tr>
<td>man</td>
<td>angry</td>
<td>2.47</td>
<td>15</td>
<td>.834</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>3.05</td>
<td>58</td>
<td>.963</td>
</tr>
</tbody>
</table>

The angry child stimuli were rated as having the highest affect magnitude, while angry woman and angry man had the lowest magnitude ratings. Sad woman and sad man were rated as having higher affect magnitudes than sad child.

### TABLE III. NATURALNESS RATING BY SPEAKER.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Emotion</th>
<th>Naturalness</th>
<th>N</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>no emotion</td>
<td>3.96</td>
<td>53</td>
<td>.999</td>
</tr>
<tr>
<td>child</td>
<td>angry</td>
<td>2.75</td>
<td>24</td>
<td>.608</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>3.47</td>
<td>19</td>
<td>.841</td>
</tr>
<tr>
<td></td>
<td>no emotion</td>
<td>3.05</td>
<td>20</td>
<td>.605</td>
</tr>
<tr>
<td>woman</td>
<td>angry</td>
<td>3.00</td>
<td>11</td>
<td>.632</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>3.13</td>
<td>23</td>
<td>.968</td>
</tr>
<tr>
<td></td>
<td>no emotion</td>
<td>3.40</td>
<td>10</td>
<td>.516</td>
</tr>
<tr>
<td>man</td>
<td>angry</td>
<td>3.93</td>
<td>15</td>
<td>.594</td>
</tr>
<tr>
<td></td>
<td>sad</td>
<td>3.81</td>
<td>58</td>
<td>.805</td>
</tr>
<tr>
<td></td>
<td>no emotion</td>
<td>3.74</td>
<td>38</td>
<td>.921</td>
</tr>
</tbody>
</table>

### IV. DISCUSSION

This experiment explored the use of the GlórCáil-DNN system and the extent to which simple global manipulations to source and filter parameters in this system can transform a baseline speaker and the perceived affect. It was envisaged as a first step towards enabling voice control within our ABAIR TTS voices for the Irish dialects. The scaling factors applied were obtained in an earlier manipulation task [3].

Overall, the results show that the GlórCáil-DNN system works as intended, and that the global manipulations do broadly achieve speaker and affect shifts in the expected direction. However, the targeted age/gender and affect shifts are achieved only up to a point: the less successful results provide useful pointers as to future directions in controlling voice variation in our TTS voices.

#### A. Speaker Identity

In this experiment, the child and woman voices were poorly differentiated. This was not surprising, given the similarity of the scaling factors used. It suggests that the problem lies not so much with the system, or even necessarily with the global nature of the manipulations used, but rather with the original data from which they are drawn [3], where the woman and child scaling factors are clearly not optimal. This points to the need for further analytic data to optimize the scaling for future experiments. Informal tests with our system suggest that transforming a female voice to male and child voices seems to produce outputs that are more readily identified ‘correctly’ in terms of target age and gender. This is something we propose to investigate.

A further finding is that male-to-female or male-to-child voice transformation do not necessarily yield optimal results in terms of naturalness when a simple single coefficient warping factor is applied to vocal tract filter coefficients. As mentioned earlier, a previous study [6] has shown that simple formant scaling can produce convincing male-to-female transformation, and the differences in the results here may lie in the different methodologies used to model the vocal tract.

#### B. Affect

Sad voice was identified correctly in about 65% of the cases (all speakers combined), but the identification of no
emotion was much lower at 45%, and angry lower again at 32%. This relatively poor signaling, particularly of the angry affect is likely to reflect the absence of utterance-internal dynamic variation of voice source parameters. Earlier studies such as [29] and [5] have suggested that the signaling of anger entails considerable dynamic variation, which is coupled to the prosodic structure of the utterance. A next step will be to look at the role of dynamic, utterance-internal modulation, exploring which affects most crucially entail such dynamic variation, and the likely linkage of this variation to the prosodic structure of utterances.

Of the three affective states, the identification rate of sad is the highest, and the no emotion stimuli were very often identified as sad. This finding suggests that the baseline voice used in this experiment may have had a voice quality (laxer voice) that predisposes listeners towards the perception of the sad affect. Future work will need to factor in any such potential bias in the baseline voice.

The baseline system was found to produce the most natural utterances, closely followed by the male voice in all affective states produced by the GlórCáil-DNN system. The woman and child voices were rated as less natural. The angry child stimuli were rated as the least natural. As suggested above, the loss of naturalness for the non-male targets is likely linked to the high degree of vocal tract warping applied, and this is also something that will need to be revisited.

V. CONCLUSIONS

This paper presents the GlórCáil analysis-synthesis system and its integration into a DNN-TTS system with a view to providing voice transformation and affect modulation in our Irish ABAIR TTS voices. A specific experiment looked at the extent to which uniform, global scaling of a set of source and filter parameters can achieve targeted age/gender and affect changes.

The results demonstrate that such global scaling can alter the perceived speaker identity and affect. However, some of the results (e.g., for anger) point to the need for utterance-internal dynamic variation of source parameters to effectively capture particular affects. Furthermore, the baseline voice quality used may bias toward affective interpretation. These are aspects that will need to be explored in future research.

REFERENCES


