Prediction of Glottal LF Parameters using Regression Trees

Michelle Tooher, John G. McKenna

School of Computing
Dublin City University, Ireland
mtooher@computing.dcu.ie
john.mckenna@computing.dcu.ie

Abstract
The behaviour of the glottal Liljencrants-Fant (LF) parameters were studied across vowel, context, duration, stress and fundamental frequency. Using statistical analysis we attempted to account for the variation in the glottal parameters for a speaker and from the patterns observed create a model capable of predicting the parameters for a given utterance and fundamental frequency. The parameters varied, as expected, with fundamental frequency, and both prosodic and contextual information were statistically significant predictors. Regression trees were used to create models for each of the parameters and predictions of the LF parameters were calculated. The average relative percentage error of the predictions varied across parameters, and the study indicated that it is possible to make acceptable predictions of the LF parameters.

1. Introduction
Conventional approaches to prosodic modelling and manipulation in speech synthesis achieve reasonable results, but little attention is paid to the glottal source. The need for an improved understanding of the voice source and how it varies in speech has previously been expressed [1], as has the need for adequate voice source control rules in an effort to completely control prosody, fundamental frequency (F0) and voice quality in speech synthesis[2].

Studies have been reported where voice quality and the voice source are analysed in vowels within varying contexts [3], however little attention has been paid to the role of a varying F0.

The work presented here attempts to predict the variation in the Liljencrants-Fant(LF) parameters of the glottal source, focusing on the effects of F0 and contextual information. It follows from a previous study in which the glottal LF parameters were calculated. The average relative percentage error of the predictions varied across parameters, and the study indicated that it is possible to make acceptable predictions of the LF parameters.

2. Method

2.1. Data
The data used is sourced from the Boston University Radio News Corpus [5]. The corpus consists of seven newscasters recorded both in the newsroom and in a speech laboratory. For the purpose of this study we used only that data recorded in the laboratory for one of the male speakers (M1B). The laboratory-recorded news stories consist of four news stories which the presenter has himself written. The stories are read in the speaker’s professional radio style. The data yielded 2818 occurrences of 17 vowels.

The speech files were automatically aligned using both the word and phoneme level transcriptions provided with the corpus, and were hand corrected. Vowels were extracted from the speech and automatically inverse filtered using a Kalman-Filter based, linear prediction technique [6]. This method of linear prediction automatically detects the location of closed-phase sections over which to perform analysis, with subsequent inverse filtering to obtain the glottal source. In cases where the above method encountered problems with inverse filtering, the problematic data was omitted.

The widely used LF model [7] of the glottal source was applied to the data, after the method described in [8], with some minor modifications.

2.2. Analysis Of Glottal Source Parameters
A previous study [4] on a small corpus of vowels in a controlled contextual and prosodic environment displayed evidence of non-linearity between the LF parameters \( t_{BC}(\text{point of maximum airflow through the glottis}) \), \( t_{CC}(\text{point of maximum negative amplitude}) \), and \( t_{CP}(\text{point of glottal closure}) \) as they vary with \( T_{0}(\text{the pitch period}) \). \( T_{0}(\text{pitch period}) \) remained relatively constant as \( T_{0} \) varied.

2.2.1. Statistical Analysis
The LF parameters were linearly regressed against \( T_{0} \) as a first step in determining the manner of variation. As reported in the previous study [4] the values of three of the timing parameters increased with \( T_{0} \), see Fig. 1, however the relationship appeared to be linear, as opposed to the non-linear trends reported in [4].
Outliers in the data generally occur at the highest and lowest of 0.11 - 0.005. In Figure 1, it can be seen that most of the vowel, context classes, duration, and stress.

As expected with such a spread uniformly over the F0 ranges of 90Hz - 210Hz (value of 0.11 - 0.005). In Figure 1, it can be seen that most of the outliers in the data generally occur at the highest and lowest values.

Linear regression was performed on each of the parameters against $T_0$, and its effect on $t_p$, $t_{v}$, $t_{c}$, and $T_{\alpha}$ was significant (see Table 1). The $R^2$ value for regression on $T_{\alpha}$ was very small, but $T_0$ was still considered a statistically significant predictor. As expected with such $R^2$ values, there appeared to be a large standard deviation from the regression line in the cases of all four parameters.

Table 1: $R^2$ values for the parameters when regressed on $T_0$, vowel, context classes, duration, and stress.

<table>
<thead>
<tr>
<th></th>
<th>$t_p$</th>
<th>$t_{v}$</th>
<th>$t_{c}$</th>
<th>$T_{\alpha}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_0$</td>
<td>.511</td>
<td>.630</td>
<td>.011</td>
<td></td>
</tr>
<tr>
<td>$T_0$ &amp; vowel</td>
<td>.521</td>
<td>.635</td>
<td>.017</td>
<td></td>
</tr>
<tr>
<td>Stepwise regression</td>
<td>.568</td>
<td>.648</td>
<td>.127</td>
<td></td>
</tr>
</tbody>
</table>

These observations clearly show there are phenomena other than $T_0$ affecting the glottal source. We performed stepwise regression on all contextual and prosodic information available. This method of regression iteratively adds and removes variables, discarding those deemed insignificant. The contextual information included the phonological features of five phones immediately preceding, and five phones following the vowel and a variable identifying the vowel itself. The prosodic information included duration, whether the vowel was stressed or unstressed (binary), $T_0$ and the relative position of the pitch period in the vowel.\(^1\)

The results of the stepwise regression showed $T_0$ as the most significant predictor of the glottal parameters ($R^2$ values as in row 1 of Table 1). The addition of other variables did improve the $R^2$ value of the model particularly in the case of $T_{\alpha}$.

Due to the linear nature of the parameters with $T_0$, the effects of contextual and prosodic information, regression trees were applied to the data in an effort to predict the parameters.

2.2.2. Prediction Of Parameters Using Regression Trees

GUIDE [9], a regression tree algorithm developed at the University of Wisconsin-Madison, was used for building the trees, which are piecewise multiple linear (linear or polynomial) regression models. These models are constructed by recursively partitioning the data. Models are fitted by maximum likelihood to each node, and the signs of the residuals determine whether data is split to the left or right side of the tree. The trees are pruned using the cost-complexity pruning method of CART [10], whereby an overly large tree is created and then sequentially pruned back until only the root node is left. This results in a sequence of nested subtrees. The tree is chosen by estimating the prediction mean squared error of each tree using N-fold cross validation, where N is the size of the learning sample.

All vowels were entered as categorical variables, their phonetic symbol representing a category. Context classes and stress were valued at 0 or 1. These variables and duration were used for splitting the nodes. $T_0$ and position were used as regressors in the models. The program was run once for each variable $t_p$, $t_{v}$, $t_{c}$, and $T_{\alpha}$ yielding a regression tree and corresponding regression coefficients for each parameter. An example of a tree produced for $t_p$ using only the vowel as a categorical variable and $T_0$ as a regressor is shown in Fig. 2.

Figure 2: $t_p$ regressed on $T_0$, using vowel as a splitting variable. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. Number in italics beneath a terminal node is the sample $t_p$-mean. $S_i$ refers to sets of vowels, eg. $S_2 = \{/AX/, /AXR/, /ER/, /OW/, /UH/\}$. A full list of vowel sets may be seen in [12]. ARPAbet was used for phonetic transcriptions.

80% of the data set was used for training the regression trees and the remaining 20% was used for testing the model. The regression coefficients derived for the trees were used to predict new LF glottal parameters.

\(^1\)Relative position refers to the number of the pitch period over the total number of pitch periods for a particular utterance. For example, 1/10 is the first of ten pitch periods.
The percentage relative error of each of the predictions for six experiments are outlined in Table 2 below and the regression coefficients calculated for Experiment 1 and Experiment 2 are shown in Table 3. The percentage error is calculated using (1), where $N$ is the number of samples.

$$\text{Avg. Relative \% Error} = \frac{\sum (\text{actual} - \text{predicted})}{\text{actual}} \times 100 \quad (1)$$

The six experiments were:

- Exp. 1: A simple regression line was calculated for all three parameters $t_p$, $t_e$, $t_c$. A constant line was calculated for $T_0$ based on the average $T_0$ value in the learning set.

- Exp. 2: Each parameter was regressed against $T_0$, yielding separate regression lines for each.

- Exp. 3: As Exp. 2, using the vowel as a splitting variable.

- Exp. 4: As Exp. 3, including position as a regressor variable.

- Exp. 5: All contextual and prosodic information was included, using $T_0$ and relative position as regressor variables and all others as splitting variables.

- Exp. 6: The data was regressed onto a quadratic, using $T_0$ as the regressor and all other information as splitting variables.

Table 2: Relative \% Errors of parameter predictions (M1B).

<table>
<thead>
<tr>
<th></th>
<th>$t_p$</th>
<th>$t_e$</th>
<th>$t_c$</th>
<th>$T_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>60.97</td>
<td>23.65</td>
<td>19.72</td>
<td>54.10</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>35.83</td>
<td>23.23</td>
<td>12.39</td>
<td>47.07</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>34.99</td>
<td>22.79</td>
<td>12.16</td>
<td>47.94</td>
</tr>
<tr>
<td>Exp. 4</td>
<td>35.21</td>
<td>22.89</td>
<td>12.21</td>
<td>46.70</td>
</tr>
<tr>
<td>Exp. 5</td>
<td>34.54</td>
<td>22.65</td>
<td>12.03</td>
<td>48.22</td>
</tr>
<tr>
<td>Exp. 6</td>
<td>31.06</td>
<td>21.48</td>
<td>11.84</td>
<td>46.56</td>
</tr>
</tbody>
</table>

As an additional study we collected data from a second male speaker (M2B) in the news corpus. The data was processed in the same manner as that in Section 2.1, but hand correction was not performed on the automatic alignment. On completion of inverse filtering and fitting of the LF model, there were 1432 occurrences of the vowels.

We created regression models for the new data, using 80% of the data for learning and 20% for testing the models. The same six experiments were performed and results are detailed in Table 4.

Table 4: Relative \% Errors of parameter predictions for M2B.

<table>
<thead>
<tr>
<th></th>
<th>$t_p$</th>
<th>$t_e$</th>
<th>$t_c$</th>
<th>$T_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>73.11</td>
<td>23.56</td>
<td>22.06</td>
<td>97.11</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>34.02</td>
<td>20.11</td>
<td>11.57</td>
<td>97.46</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>32.10</td>
<td>18.95</td>
<td>11.34</td>
<td>96.81</td>
</tr>
<tr>
<td>Exp. 4</td>
<td>32.16</td>
<td>19.17</td>
<td>11.53</td>
<td>96.81</td>
</tr>
<tr>
<td>Exp. 5</td>
<td>35.72</td>
<td>24.24</td>
<td>11.74</td>
<td>97.21</td>
</tr>
<tr>
<td>Exp. 6</td>
<td>31.02</td>
<td>18.87</td>
<td>10.92</td>
<td>94.12</td>
</tr>
</tbody>
</table>

One extra set of experiments was also performed to test speaker dependence of the prediction models created for the first male speaker (M1B). The test data for M2B was applied to the regression models obtained for M1B, and results are reported in Table 5.

Table 5: Relative \% Errors of parameter predictions, using the models derived for M1B and testing them with the test data from M2B.

<table>
<thead>
<tr>
<th></th>
<th>$t_p$</th>
<th>$t_e$</th>
<th>$t_c$</th>
<th>$T_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>85.50</td>
<td>28.58</td>
<td>17.69</td>
<td>97.11</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>48.46</td>
<td>28.02</td>
<td>15.02</td>
<td>98.84</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>46.12</td>
<td>27.51</td>
<td>14.85</td>
<td>98.85</td>
</tr>
<tr>
<td>Exp. 4</td>
<td>46.27</td>
<td>27.41</td>
<td>14.84</td>
<td>98.49</td>
</tr>
<tr>
<td>Exp. 5</td>
<td>52.01</td>
<td>35.71</td>
<td>16.97</td>
<td>129.27</td>
</tr>
<tr>
<td>Exp. 6</td>
<td>39.49</td>
<td>26.37</td>
<td>14.10</td>
<td>84.50</td>
</tr>
</tbody>
</table>

3. Discussion

The relative percentage errors of the LF parameter predictions (in Table 2) are encouraging, suggesting that it is possible to obtain reasonable estimates of the LF parameters given utterance specific information and it is possible that with additional information about the utterances (eg. style, emotion) we could improve these results further. The improvement in the error across experiments is, however, quite small and it is possible that the computational effort involved by adding more information might overshadow the improvement in results.
We have presented a study investigating prosodic and contextual influences on the glottal parameters for a speaker; both influence the LF glottal timings and the results indicate that they are statistically significant predictors of the glottal parameters. The parameters were seen to vary differently across $T_0$ suggesting that conventional approaches to manipulating the glottal waveform might be limited in their ability to truly reflect the changes in the glottal source. Regression trees were fitted to the data in an attempt to predict the glottal parameters, given prosodic and contextual information, yielding favourable results. A more sophisticated approach to glottal waveform manipulation taking into consideration the effects of vowel quality, prosody and phonetic environment holds the possibility of improving the naturalness of synthesised speech.

6. Acknowledgements
Michelle Tooher is partially funded by an Enterprise Ireland Scholarship Ref. No. BR/2001/158.

7. References