ENHANCED PITCH TRACKING AND THE PROCESSING OF F0 CONTOURS FOR COMPUTER AIDED INTONATION TEACHING

P.C. Bagshaw, S.M. Hiller, and M.A. Jack

Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh, EH1 1HN, Scotland, UK

ABSTRACT
A comparative evaluation of several pitch determination algorithms (PDAs) is presented. Fundamental frequency estimates, F0, are compared with largenal frequency estimates, Lx. An algorithm is presented which enables Lx contours to be generated from laryngograph data. We seek the most accurate method of F0 extraction in order to minimise errors propagating into subsequent prosodic analysis. The super resolution pitch determinator [3] performs well relative to the other PDAs studied. Modifications made to this algorithm are described, which radically reduce the number of gross F0 errors and improve the classification of voiced and unvoiced sections of speech. The raw F0 contours produced by this enhanced algorithm are processed to form schematised contours used in computer aided intonation teaching. The series of processes used in the schematisation is described.

Keywords: Pitch tracking, Intonation, Language teaching

1 INTRODUCTION
The fundamental frequency of speech plays an important role in the prosodic features of stress, rhythm, and intonation. The understanding and appropriate use of prosody is an important component of foreign language learning for both comprehension and intelligibility. Computer aided teaching of intonation therefore requires the determination of F0 as an initial process in the automated assessment of the speech of a non-native student. Determining F0 is not a simple task, and many approaches have been reported [2]. The selection of PDAs investigated here covers a range of techniques which use both time domain and frequency domain representations of speech. An evaluation of the algorithms, based on the use of laryngograph data, is described in Section 2. A method of forming a ‘reference’ contour from laryngograph data is presented. F0 contours generated by each PDA are compared with the ‘reference’ Lx contours. The evaluation shows that the super resolution pitch determinator has the potential to form accurate F0 contours from low-pass filtered speech. Errors occurring during F0 extraction must be minimised to prevent them from propagating into the prosodic analysis. Enhancements described in Section 3 are made to this algorithm in order to minimise F0 errors. The F0 contour produced by a PDA is, however, not manipulated solely by linguistic and paralinguistic effects. An F0 contour is also affected by segmental content, micro-perturbations, the speaker’s anatomy and physiology, and errors involved in its determination from the speech waveform. Such F0 variations need to be removed in order to facilitate the comparison of a student’s intonation with that of a native speaker. This is performed by a series of post-processes which schematises a raw F0 contour and is described in Section 4.

2 EVALUATION OF PITCH DETERMINATION ALGORITHMS
Seven PDAs are investigated. Their selection was influenced by availability, ease of implementation, and by the desire to examine methods of F0 extraction which use radically different techniques.

- Cepstrum pitch determination (CPD) [4].
- Feature-based pitch tracker (FBPT) [6].
- Harmonic product spectrum (HPS) [11] [5].
- Integrated pitch tracking algorithm (IPTA) [12].
- Parallel processing method (PP) [1].
- Super resolution pitch determinator (SRPD) [3].
- Enhanced version of SRPD (εSRPD).

The functionality of all of the algorithms is dependent upon certain thresholds and pre-determined parameters, some of which are common across algorithms. In order to set a degree of similarity between the PDAs, all are required to present a computed F0 value at 6.4ms intervals. The values are limited to the ranges of 50Hz–250Hz for male speakers and 120Hz–400Hz for female speakers. In cases where a fixed-length analysis frame is required by an algorithm, the frame duration is set to 38.4ms. This duration enables at least two signal periods to reside within the frame for all F0 values greater than 52Hz, and allows sufficient data for cepstral and spectral analysis techniques. The speech data is sampled at 20kHz using a 16-bit analogue-to-digital converter. Some of the PDAs require a low-pass filtered version of this data, which is produced by an FIR filter with a -3dB cut-off at 600Hz.
2.1 A Laryngeal Frequency Tracker

The ‘reference’ contours are created from laryngograph data recorded simultaneously with speech by using a simple ‘pulse’ (Fig. 1) location algorithm and deriving the duration between successive pulses.

![Fig. 1: Laryngograph 'Pulse'](image)

The pulse start time \( t_{\text{start}} \) is the first sample for which the amplitude is less than zero and less than or equal to the amplitude of following samples. The pulse stop time \( t_{\text{stop}} \) is the last sample for which the amplitude is less than zero and less than or equal to the amplitude of preceding samples. The pulse width \( t_{\text{width}} \) is defined as the difference between \( t_{\text{start}} \) and \( t_{\text{stop}} \). The pulse peak amplitude \( a_{\text{peak}} \) is the maximum amplitude of samples between \( t_{\text{start}} \) and \( t_{\text{stop}} \) (always greater than zero). The pulse instant \( t_{\text{pulse}} \) is defined as the time of the first of these samples with an amplitude \( a_{\text{peak}} \). For laryngograph data sampled at 20kHz, a pulse at \( t_{\text{pulse}} \) is classed as a marker of the glottal closure instant if the pulse width \( t_{\text{width}} \) is greater than four samples and the pulse peak amplitude \( a_{\text{peak}} \) is greater than some arbitrary threshold value. The duration between one pulse instant \( t_{\text{pulse}} \) and the next \( t_{\text{pulse}}^{+1} \) is calculated and converted to Hertz. If the value lies within a limited range, it is taken to represent the laryngeal frequency at the time \( (t_{\text{pulse}}^{+1} + t_{\text{pulse}}^{+1})/2 \); otherwise, the duration between the pulses is considered to correspond to an unvoiced region of speech. The \( Lx \) limits are \( \geq 50Hz \) for male speakers, and \( \geq 120Hz \) for female speakers. There must be at least three laryngograph pulses in each voiced section. This final restriction is imposed to remove the few errors when a ‘pulse’ in the laryngograph data is formed by events other than glottal activity.

The accuracy with which \( Lx \) can be determined by this method is limited by the time quantisation in sampling the laryngograph signal. Each value of \( Lx \) has an error of \( F_s / (F_s/Lx^2 - 1) \) Hz, where \( F_s \) is the sampling frequency.

2.2 Comparative Evaluation

A database containing approximately 5 minutes of speech was used for the evaluation. It was formed from sentences read by one male and one female, and was biased towards utterances containing voiced fricatives, nasals, liquids and glides, since PDAs generally find these difficult to analyse. The quantisation error in determining \( Lx \) has a mean of 0.80Hz and population standard deviation of 0.34Hz for the male of PDAs and 0.60Hz of males and standard deviation of 0.86Hz for the female speaker. This error cannot be compensated for and affects the evaluation results for the various PDAs shown in Table 1. Durations of unvoiced or silent regions incorrectly classified as voiced by a PDA, and durations of voiced sections erroneously classified as unvoiced, are accumulated over all the utterances in the database for each speaker, and expressed as a percentage of the total duration of unvoiced (or silent) speech and voiced speech respectively. The total number of comparisons for which the difference between \( F0 \) and the reference \( Lx \) is higher or lower than 20% of \( Lx \) (gross errors) are expressed as a percentage of the total number of comparisons for which \( F0 \) and \( Lx \) represent voiced speech. The population standard deviation (p.s.d.) and the mean, absolute deviation of the \( Lx \) and \( F0 \) contours are given for when both represent voiced speech, and the PDA has not made a gross error.

<table>
<thead>
<tr>
<th>PDA</th>
<th>Unvoiced in error (%)</th>
<th>Voiced in error (%)</th>
<th>Gross errors High (%)</th>
<th>Gross errors Low (%)</th>
<th>Absolute deviation (Hz) mean</th>
<th>p.s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPD</td>
<td>18.11</td>
<td>19.89</td>
<td>4.00</td>
<td>0.64</td>
<td>2.34</td>
<td>3.06</td>
</tr>
<tr>
<td>FRPT</td>
<td>3.73</td>
<td>13.90</td>
<td>1.27</td>
<td>0.64</td>
<td>1.86</td>
<td>2.89</td>
</tr>
<tr>
<td>HPS</td>
<td>14.11</td>
<td>7.07</td>
<td>5.34</td>
<td>28.15</td>
<td>3.25</td>
<td>3.21</td>
</tr>
<tr>
<td>IPTA</td>
<td>9.78</td>
<td>17.4</td>
<td>1.40</td>
<td>0.83</td>
<td>2.67</td>
<td>3.37</td>
</tr>
<tr>
<td>PP</td>
<td>7.69</td>
<td>15.82</td>
<td>0.22</td>
<td>1.74</td>
<td>2.64</td>
<td>3.01</td>
</tr>
<tr>
<td>SRPD</td>
<td>4.05</td>
<td>16.78</td>
<td>0.02</td>
<td>2.01</td>
<td>1.78</td>
<td>2.46</td>
</tr>
<tr>
<td>CRPD</td>
<td>4.63</td>
<td>12.04</td>
<td>0.90</td>
<td>0.56</td>
<td>1.40</td>
<td>1.74</td>
</tr>
<tr>
<td>CRPD</td>
<td>31.63</td>
<td>22.22</td>
<td>0.61</td>
<td>3.94</td>
<td>6.39</td>
<td>7.01</td>
</tr>
<tr>
<td>FBPT</td>
<td>3.61</td>
<td>12.16</td>
<td>0.60</td>
<td>3.55</td>
<td>5.46</td>
<td>7.03</td>
</tr>
<tr>
<td>HPS</td>
<td>19.10</td>
<td>21.06</td>
<td>0.46</td>
<td>1.01</td>
<td>4.59</td>
<td>5.31</td>
</tr>
<tr>
<td>IPTA</td>
<td>5.70</td>
<td>15.93</td>
<td>0.53</td>
<td>3.12</td>
<td>4.38</td>
<td>5.35</td>
</tr>
<tr>
<td>PP</td>
<td>6.15</td>
<td>13.01</td>
<td>0.26</td>
<td>3.20</td>
<td>6.11</td>
<td>6.45</td>
</tr>
<tr>
<td>SRPD</td>
<td>2.35</td>
<td>12.16</td>
<td>0.39</td>
<td>5.56</td>
<td>4.14</td>
<td>5.51</td>
</tr>
<tr>
<td>CRPD</td>
<td>2.35</td>
<td>9.13</td>
<td>0.48</td>
<td>0.23</td>
<td>1.17</td>
<td>5.13</td>
</tr>
</tbody>
</table>

Table 1: PDA evaluation for male speech (top) and female speech (bottom)

Any PDA producing an \( F0 \) contour suitable for intonation analysis must perform consistently between male and female speech. The resultant contour must accurately determine voicing so that pitch accents are not left undetected and gross \( F0 \) errors must be minimal for subsequent processing. CPD and HPS are therefore unsuitable algorithms for the task in hand. Of those remaining, FBPT and SRPD form the best voiced/unvoiced classifications. SRPD would be the most suitable algorithm if the voiced/unvoiced classification performance could be improved and the number of gross (\( F0 \) too low) errors reduced.

3 ENHANCED SUPER RESOLUTION PITCH DETERMINATOR (eSRPD)

The speech is initially low-pass filtered. Frames of data for which an \( F0 \) estimate is required at the time of sample \( s(i) \) are divided into three consecutive segments of \( n \) samples,

\[ x_n = \{ x(i) = s(i - n) \mid i = 1 \text{ to } n \} \]
\[ y_n = \{ y(i) = s(i) \mid i = 1 \text{ to } n \} \]
\[ z_n = \{ z(i) = s(i + n) \mid i = 1 \text{ to } n \} \]

Frames at the beginning of an utterance, for which \( x_n \) is not fully defined, are classified as 'silent'; likewise frames
The value of $n$ is optimised so that each segment occupies a fundamental period. The optimisation selects a value of $n$ within a limited range, $N_{\text{min}}$ to $N_{\text{max}}$ samples, which is directly related to the expected range of $F0$ values for a given speaker. The minimum and maximum values of the sample sets $x_{N_{\text{min}}}$ and $y_{N_{\text{max}}}$ are determined. If the sum of these (absolute) values is less than some preset threshold for either set, then the frame is classified as ‘silent’. Otherwise, the coefficient $p_{x,y}(n)$ is determined for the values of $n$ within the limited range in steps of a decimation factor $L$ ($L = 4$ in this investigation).

$$p_{x,y}(n) = \frac{\sum_{j=1}^{[n/L]} x(jL), y(jL)}{\sum_{j=1}^{[n/L]} x(jL)^2, \sum_{j=1}^{[n/L]} y(jL)^2}$$  \hspace{1cm} (2)

where $n = N_{\text{min}} + iL \, \mid \, i = 0, 1, \ldots; N_{\text{min}} \leq n \leq N_{\text{max}}$

$p_{x,y}(n)$ is invalid if the number of zero-crossings in $x_n + y_n$ is less than 4. The locations of local maxima in $p_{x,y}(n)$ with values above an adaptive threshold (as described by Medan et al. [3]) form candidates for the optimum value of $n$. If no candidates for the fundamental period are found, the frame is classified as ‘unvoiced’. Otherwise, the frame consists of ‘voiced’ speech, and a second coefficient, $p_{y,z}(n)$ is determined for all the fundamental period candidates. Those candidates for which $p_{y,z}(n)$ also exceeds the threshold value are given a score of 2, while the others are given a score of 1. Candidates with a higher score are more likely to represent the true fundamental period. If there are one or more candidates with a score of 2, then all those with a score of only 1 are removed from the list of candidates and ignored. Following this, if there is only one candidate (with a score of either 1 or 2) the candidate is assumed to be the best estimate of the fundamental period for that frame. Otherwise, the candidates are listed in order of increasing fundamental period. The candidate at the end of this list is selected to represent a fundamental period of $n_M$, and the $m$th candidate a period $n_M$. Another coefficient, $q(n_M)$, is calculated for each candidate, where $q(n_M)$ is the correlation coefficient between two segments of length $n_M$ spaced $n_M$ apart.

$$q(n_M) = \frac{\sum_{j=1}^{n_M} s(j - n_M), s(j + n_M)}{\sum_{j=1}^{n_M} s(j - n_M)^2, \sum_{j=1}^{n_M} s(j + n_M)^2}$$  \hspace{1cm} (3)

The first coefficient $q(n_1)$ is then assumed to be the ideal. If a subsequent $q(n_M)$ exceeds this ideal when multiplied by 0.77 (arbitrary) then it is in turn assumed to be the new ideal. The candidate for which the value of $q(n_M)$ is believed to be ideal is taken as the best estimate for the fundamental period of the frame being analysed.

Biasing is applied to the coefficients $p_{x,y}(n)$ and $p_{y,z}(n)$ for values of $n$ where the fundamental period of a new frame is expected to lie, if the two previously analysed frames were classified as ‘voiced’; if the $F0$ value of the previous frame is not being temporarily held, and if the fundamental frequency of the previous frame $f_0^{\text{prev}}$ is less than $\frac{2}{3}$ times the fundamental frequency of its preceding voiced frame $f_0^{\text{prev}}$, and greater than $\frac{5}{8} f_0^{\text{prev}}$, i.e. if it is highly probable that the fundamental period estimate of the previous frame is not erroneous. The fundamental period of the new frame $n_0^{\text{new}}$ is expected to lie within the range of $n$ closest to $n_0^{\text{prev}}$ for which the set of $p_{x,y}(n)$ from the previous frame are greater than zero (Fig. 2). The coefficients $p_{x,y}(n)$ and $p_{y,z}(n)$ are doubled for values of $n$ in this range. This effectively applies a bias on the location of a maxima in the region of the fundamental period for the previous frame to form a candidate for the fundamental period of the current frame. Note, however, that the voiced/unvoiced decision is based on the presence or absence of local maxima in $p_{x,y}(n)$ which exceed the adaptive threshold. The biasing will therefore tend to increase the percentage of unvoiced regions being incorrectly classified as ‘voiced’. In order to minimise this undesirable side effect, if the unbiased coefficient $p_{x,y}(n)$ does not exceed the threshold for the candidate believed to be the best estimate of the frame fundamental period, then the $F0$ value for that frame is held until the state of the subsequent frame is known. If this next frame is classified as ‘silent’ or ‘unvoiced’, the former frame is re-classified as ‘unvoiced’.

![Fig. 2: Example Set of $p_{x,y}(n)$ from Previous Frame](image)

The remainder of the processing to refine the accuracy of the $F0$ estimate is as described by Medan et al. [3]. The results shown in Table 1 for the enhanced SRPD algorithm ($e$SRPD) demonstrate the effects of these modifications on
4 POST-PROCESSING OF F0

The number of short sections of gross F0 errors that occur during F0 extraction are reduced by applying a non-linear smoother [7] spanning 17 frames (with a 6.4ms interval between frames). Linear smoothing, using a hanning window, is also applied to remove small perturbations in the contour, and reduce the effect of any remaining quantisation errors. A window of length l takes l consecutive values and weights the n\textsuperscript{th} value by a factor h(n). The output of the linear smoother is the sum of the weighted values

\[ h(n) = \frac{1}{l+1} \left( 1 - \cos \frac{2\pi n}{l+1} \right) \text{ for } 1 \leq n \leq l \quad (4) \]

A smoothed F0 contour may exhibit an overall downward trend in F0 during the course of an utterance. Such declination may result in two accents which have the same perceived pitch having different F0 values. Compensation for this effect is attempted by initially applying least mean of squares regression [10] to all the local maxima and to all the local minima in the F0 contour. An average line is taken between the two resultant linear models and used as an estimate of the declination. Declination-compensation is only applied if this average line has a negative slope. The mean and the population standard deviation of the pre-declination-compensated F0 contour are retained by using frequency shifting and scaling.

Z-score normalisation [9] is applied to the declination-compensated contour to enable different F0 values from different speakers to correspond to the same phonological pitch. An observed F0 value is expressed as a multiple of a measure of dispersion relative to the mean F0. The normalised F0 value is given by,

\[ F0_{\text{norm}} = \frac{F0_{\text{input}} - \overline{F0}}{\sigma} \quad (5) \]

where \( \overline{F0} \) is the long-term mean F0 for a given speaker, and \( \sigma \) is the long-term population standard deviation.

Short breaks in continuity of the normalised contour are filled by linear interpolation. Breaks are only filled if they have a duration of less than 80ms and if the jump in \( F0_{\text{norm}} \) across the break is less than \( \sigma/2 \). The contour is then smoothed again through a hanning window spanning 17 frames.

This series of processes forms a schematicised F0 contour which is used in intonation analysis for foreign language teaching [8].

5 CONCLUSIONS

Laryngograph signals, which are recorded simultaneously with speech, have been used to derive ‘reference’ Lx contours. These have been used to evaluate a selection of pitch determination algorithms. An enhanced super resolution pitch detector (eSRP) is used to offer improved performance relative to the other PDAs reported in this study, with less than 1.5\% gross F0 errors and less than 16.7\% of speech classified as voiced or unvoiced incorrectly. The new eSRP algorithm has been shown to perform well independently of a speaker’s sex and is unlikely to leave pitch accents undetected. This PDA is therefore the most suitable of those investigated for applications in computer aided intonation teaching where a raw F0 contour is smoothed, compensated for the declination of F0 over an utterance, normalised for speaker differences in long-term F0 level and range, interpolated over small gaps, and smoothed again, to form a schematicised F0 contour. The schematicised contour is used as one of the main inputs to an automatic system for intonation teaching.

REFERENCES


