Reliable tracking based on speech sample salience of vocal cycle length perturbations

C. Mertens 1, F. Grenez 1, L. Crevier-Buchman 2 J. Schoentgen 1,3

1 Laboratory of Images, Signals and Telecommunication Devices, Faculte des Sciences Appliquees, Universite Libre de Bruxelles, Brussels, Belgium
2 Hopital Europeen Georges Pompidou, Paris, France
3 National Fund for Scientific Research, Belgium

chmerten@ulb.ac.be, fgrenez@ulb.ac.be, lise.buchman@numericable.fr, jschoent@ulb.ac.be

Abstract

The presentation concerns a method for tracking cycle lengths in voiced speech. The speech cycles are detected via the saliences of the speech signal samples, defined by the length of the temporal interval over which a sample is a maximum. The tracking of the cycle lengths is based on a dynamic programming algorithm which does not request that the signal is locally periodic and the average period length known a priori. The method has been validated on a corpus of normophonic speakers. The results report the tremor frequency and modulation depth of the vocal frequency of 72 ALS and 8 normophonic speakers.

Index Terms: vocal frequency, vocal tremor, speech sample salience analysis, dynamic programming

1. Introduction

In clinical applications of speech analysis, speech cycles are detected to measure their lengths and amplitudes with a view to investigating slow (vocal tremor) and fast (vocal jitter and shimmer) perturbations of vocal frequency and speech cycle amplitude. Often, such analyses are frame-based and the cycle detection rests on the recursive detection and storage of speech signal extrema that occur in the vicinity of the instants of glottal excitation. To enable this selection, one often assumes that voiced speech segments are pseudo-periodic so that the peaks can be selected one by one on the base of a prior estimation of the typical fundamental period. The assumption of quasi-equal spacing is, however, valid for modal voices only and not for pathological ones, which may be characterized by large cycle-to-cycle fluctuations in length or amplitude. Cycle insertion or omission errors may therefore occur, which bias the acoustic cues of cycle regularity.

Here, we propose to track speech cycles via a multi-scale analysis that assigns a salience to each signal peak. The speech cycle tracking is founded on a dynamic programming paradigm. It does not rest on the assumptions that the speech signal is locally periodic and the average period length known a priori. A signal peak is a signal sample whose left and right neighbours are smaller. The salience of a speech signal peak designates the time interval over which this peak is a maximum.

Several methods for the tracking of cycle lengths by means of saliences have been considered before. In [1], the cycle detection was based on the selection of a single peak sequence in order of decreasing salience value. In [2], several candidate lengths sequences were generated and the most regular retained.

A major drawback of both proposals was the lack of a constraint on the regularity of the cycle lengths sequence, which therefore was subject to spurious perturbations in the case of salience reaffiliation between primary and secondary peaks in the same cycle, for instance. Another disadvantage was that the selection of each cycle marker was final without the option to drop it in favor of a more likely candidate yet to be discovered.

Therefore, here we propose a vocal cycle detection method that relies on dynamic programming to extract a cycle sequence the length perturbations of which is minimal. The cost function involves the second order differences of successive speech cycle durations as well as the cycle peak saliences. The tracker does not rely on estimates of the typical cycle length, as opposed to existing proposals involving dynamic programming in the extraction of the vocal frequency or glottal cycle length [3].

Section 2 explains speech sample saliences and the tracking of the speech cycles, as well as the corpora. Section 3 reports the results and discussion of the tracking of slow perturbations in ALS and normophonic speakers.

2. Method

2.1. Preprocessing

The speech signal has been band-pass filtered by means of a finite response (FIR) filter with cut-off frequencies equal to 60 Hz and 1000 Hz to remove additive low-frequency hum, additve noise owing to turbulence as well as high-frequency formants.

The speech signal has then been upsampled to for to remove additive low-frequency hum, and additive noise owing to turbulence as well as high-frequency formants. The signal peak is then the peak position to be measured with a precision requested by the size of vocal jitter, which in modal voice is expected to be < 1% of the typical cycle length.

2.2. Speech sample salience analysis

The salience of a signal sample (which may be a signal peak or not) is defined as the length of the longest temporal interval over which the signal sample is a maximum. A property of the salience is that a sample with a large salience has not necessarily a large amplitude and vice versa. In voiced speech segments, for instance, speech cycles are often characterized by a prominent signal peak that is the effect of the glottal excitation. The salience of that peak is expected to be high irrespective of the evolving signal amplitude.

The basic salience analysis algorithm is the following [1].

1. The initialization consists in assigning a salience of one
to each sample in an array of length $M$.

2. The array is subdivided into non-overlapping intervals of length $n$, starting from $n = 2$. The length of the rightmost interval is therefore comprised between 1 and $n$.

3. For each interval, the sample with the highest amplitude is detected and a salience equal to the interval length assigned to that maximal sample.

4. The analysis interval length is increased by one and steps 2. to 4. are repeated till $n = M$.

Problems with this basic salience allocation algorithm are that, generally speaking, the position of the analysis interval with regard to the signal array boundaries may bias the sample saliences, which is a known difficulty in multi-scale analysis, of which salience analysis is an instance. Known solutions such as rotating the samples within the analysis array and repeating the salience analysis has been observed to increase processing time considerably. To speed up processing and alleviate analysis interval boundary effects, partial salience analysis based on a sliding analysis window of length $N$ has been implemented.

The windowed salience analysis involves speeding up the algorithm by computing left and right-hand saliences and computing saliences for a subset of samples only. Details are reported in [1]. Other algorithmic steps are summarized hereafter.

1. The initial position of the analysis window is at the left edge of the analysis array of length $M$.

2. For each window position the saliences are updated according to the speeded-up procedure.

3. The window is shifted by one sample and steps 2. and 3. are repeated till the rightmost edge of the window coincides with the rightmost boundary of the sample array.

The final speech sample saliences are comprised between 1 and $2N - 1$. It is recommended to discard the $N - 1$ first and last sample saliences, because they are conditioned by the array boundaries. The sliding analysis window length must therefore be chosen so as to minimize the loss of information owing to the array boundaries and maximize the relevance of the window-determined saliences with regard to the goal of the multi-scale analysis.

It is easy to show that this algorithm yields salience values identical to the "true" (re basic algorithm) values as long as the distance between a signal sample and its neighbour, the amplitude of which is higher, is $< N$ (i.e. the analysis window length). If this is not the case, the saliences are clipped to one of three values, which are $2N - 1$ or $N$ + the "true" right salience or $N$ + the "true" left salience.

![Diagram](image)

Fig. 1 illustrates the peak salience values obtained for a fragment of vowel [a]. One observes that the prominent signal peaks due to the glottal excitation have a higher salience value than other secondary peaks that are due to tract resonances.

2.3. Speech cycle tracking

For speech cycle tracking, no strong assumptions are made with regard to the regularity of the cycle lengths. One assumes that the vocal frequency is comprised between $F_{\text{min}} = 60Hz$ and $F_{\text{max}} = 400Hz$, which corresponds to an inter-cycle peak spacing in samples comprised between $d_{\text{min}}$ and $d_{\text{max}}$.

The first stage consists in ranking the signal peaks according to decreasing salience and keeping those peaks the salience values of which are greater than or equal to 150% the length of the shortest possible cycle. The initial number of peaks is therefore in excess of the number of expected cycles because a typical salience value of a speech cycle peak is equal to twice the cycle length.

The second stage consists in considering several candidate cycle length time series obtained by means of the retained peak distances and discovering via dynamic programming the length series that has the smallest overall cycle duration perturbation. The candidate cycle length series are built by taking into account several signal peak sub-sequences on the base of the local inter-peak durations and the peak salience values, assuming that prominent speech cycle peaks owing to the glottal excitation are characterized by large salience values. This second stage comprises an initialization, search and backtracking step.

2.3.1. Initialization

The initialization consists in determining for each signal peak $i$ all candidate pairs of preceding peaks ($g, h$) for which the inter-peak distances satisfy the following conditions.

\[
\begin{align*}
\alpha d_{g,h} & < d_{h,i} < \beta d_{g,h} \\
\alpha d_{g,h} & < d_{h,i} < \beta d_{g,h}
\end{align*}
\]

Symbol $d_{g,h}$ designates distance in number of samples between peaks $g$ and $h$. The values of symbols $\alpha$ and $\beta$ are fixed to 0.8 and 1.2 respectively. They limit the maximal local length perturbation the tracker is able to deal with.

Several triplets ($g, h, i$) are thus obtained for each peak $i$. For each signal peak $i$ that is positioned among the first $2d_{\text{max}}$ samples in the array, two running costs are initialized and assigned to each triplet ($g', h', i'$): a cumulative cost $C_{g' h' i'}$, which is a measure of the overall cycle-to-cycle length perturbations, and the number of cycles $l_{g' h' i'}$ which are comprised in the sequence.

\[
\begin{align*}
C_{g' h' i'} &= 0 \\
l_{g' h' i'} &= 2
\end{align*}
\]

2.3.2. Optimal path search

Optimal search via dynamic programming here consists in finding the path involving those candidate cycle peaks that give rise to the smallest perturbations of the inter-peak durations. In a first step, for any triplet ($g, h, i$), several local costs are computed. Local cost (3) is based on the second order difference of the inter-peak durations and the salience of peak $i$. As a consequence, the peaks that are taken into account are the triplet

![Diagram](image)
Figure 2: Local distances

\[(g, h, i) \text{ but also all predecessor triplets } (f, g, h) \text{ which satisfy conditions (1).} \]

\[c(g, h, i)|_f = \frac{(2d_{(g,h)} - d_{(h,i)} - d_{(f,g)})^2 + 1}{s(i)} \]

(3)

In a second step, for each triplet of peaks \(g, h\) and \(i\), the optimal predecessor of \(g\) is selected as follows:

\[f^* = \arg \min_{f \in (f, g, h)} \left\{ \frac{C_{(f, g, h)} + c(g, h, i)|_f}{l_{(f, g, h)} + 1} \right\} \]

(4)

The purpose of the denominator in (4) is to disfavor cycle omissions. The two cumulative costs (5) are then updated on the basis of that optimal predecessor \(f^*\) for each triplet \((g, h, i)\) and stored in a table at the \((g, h, i)\) entry.

\[C_{(g,h,i)} = C_{(f^*, g, h)} + c(g, h, i)|_{f^*} \]

\[l_{(g,h,i)} = l_{(f^*, g, h)} + 1 \]

(5)

2.3.3. Backtracking

When all costs have been updated, the peak triplet within the last \(d_{max}\) samples giving rise to a minimal cumulative cost \(C\) divided by the number of cycles squared is kept.

The sample salience analysis and the cycle detection by dynamic programming are carried out once for each polarity of the signal. The polarity giving the smallest cumulative cost \(C\) (divided by the squared number of cycles) is retained. The peak sequence corresponding to this optimal path is then recovered by backtracking, starting from the optimal triplet \((g^*, h^*, i^*)\) and detecting the optimal preceding peak \(f^*\) via (4), and so on.

2.4. Corpora

A first corpus of normophonic speakers has comprised 8 subjects sustaining French vowels [\(a\)] for which both acoustic and contact microphone signals have been recorded simultaneously. The finite sum (integral) and difference (derivative) of the acoustic signal have also been obtained. A second corpus of dysphonic speakers has comprised 72 patients with amyotrophic lateral sclerosis (ALS), a neurological disease that affects the muscular system. The latter stimuli have been recorded at the Hôpital Européen Georges Pompidou in Paris via an acoustic microphone only.

2.5. Validation

The reliability of the tracking of the cycle length time series via speech peak saliences and regularity constraints has been tested by means of modal speech signals, their numerical derivatives and integrals as well as the co-recorded throat microphone signals (corpus 1). The four time series that are so obtained for each speaker are expected to be very similar given that they report the same glottal cyclicity via four different signals. The agreement between the four time series is evaluated by means of their inter-correlation.

In addition, the low-frequency spectra of the four time series have been obtained and inter-correlated. The reason is that in a second experiment, the tremor frequency is estimated based on the low-frequency spectrum of the cycle duration time series.

2.6. Vocal cues

In a first stage, the cycle length time series is interpolated and resampled to obtain a time series of lengths sampled at a constant sampling step. Then the average is subtracted from the time series and the discrete-time Fourier transform is computed. One acoustic cue is the abscissa of the center of gravity of the low-frequency amplitude spectrum of the cycle length series. Two frequency intervals ([0-15Hz] or [3-15Hz]) have been considered. Indeed, cardiac beat, breathing and bloodflow are expected to influence strongly the spectrum below 3Hz. The sec-
ond cue is the coefficient of variation of the cycle length series, which characterizes the excursion of the cycle durations with respect to the average. These two cues are rough estimates of respectively the modulation frequency and the modulation depth owing to the tremor and jitter of the vocal frequency.

3. Results and discussion

3.1. Experiment 1

The cycle length series have been estimated for the corpus of normophonic speakers. Visual inspection of the cycle duration time series (Fig. 3) shows that the speech cycles are correctly discovered. The signal polarity that has been retained is generally different for the four time series, but the event sequence corresponds to the instants of glottal excitation. One observes a good agreement between the four time series reporting glottal cycle durations for most of the time series. Occasionally, one observes discrepancies between time series recorded via the acoustic and throat microphones. These discrepancies are due to differences in the transducers rather than to difficulties with the signal processing.

Table 1 reports the inter-correlation coefficients for the length time series obtained from the four signals for each speaker. The corresponding inter-correlations of the low-frequency spectra have all been ≥ 0.99.

Table 1: Inter-correlation coefficients for cycle length time series obtained via the acoustic speech signal (1), its derivative (2) and integral (3) as well as the throat microphone signal (4) for each speaker

<table>
<thead>
<tr>
<th>i</th>
<th>F0</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1-2</td>
</tr>
<tr>
<td>1</td>
<td>90</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>114</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>118</td>
<td>0.95</td>
</tr>
<tr>
<td>4</td>
<td>119</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>210</td>
<td>0.98</td>
</tr>
<tr>
<td>6</td>
<td>225</td>
<td>0.99</td>
</tr>
<tr>
<td>7</td>
<td>239</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td>241</td>
<td>0.99</td>
</tr>
</tbody>
</table>

3.2. Experiment 2

The cycle duration time series have been obtained for the 72 ALS and 8 normophonic speakers. The tremor frequency has been characterized via the position of the center of gravity of the LF-spectrum of the length time series and the modulation depth by the coefficient of variation of the length time series. Table 2 reports the quartiles of the tremor frequency and modulation depth for the ALS and normophonic speakers. One observes that for the ALS speakers the modulation depth increases. The differences in tremor frequency are feeble.

Table 2: Estimates of the modulation frequency (center of gravity) and modulation depth (coefficient of variation) of vocal tremor for normophonic and ALS speakers

(a) Normophonic speakers

<table>
<thead>
<tr>
<th>Center of gravity (Hz)</th>
<th>C.V.(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 – 15Hz]</td>
<td>[15 – 30Hz]</td>
</tr>
<tr>
<td>Minimum</td>
<td>4.79</td>
</tr>
<tr>
<td>First quartile</td>
<td>5.16</td>
</tr>
<tr>
<td>Median</td>
<td>5.45</td>
</tr>
<tr>
<td>Third quartile</td>
<td>5.70</td>
</tr>
<tr>
<td>Maximum</td>
<td>6.08</td>
</tr>
</tbody>
</table>

(b) ALS speakers

<table>
<thead>
<tr>
<th>Center of gravity (Hz)</th>
<th>C.V.(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 – 15Hz]</td>
<td>[15 – 30Hz]</td>
</tr>
<tr>
<td>Minimum</td>
<td>3.18</td>
</tr>
<tr>
<td>First quartile</td>
<td>4.30</td>
</tr>
<tr>
<td>Median</td>
<td>4.88</td>
</tr>
<tr>
<td>Third quartile</td>
<td>5.69</td>
</tr>
<tr>
<td>Maximum</td>
<td>6.86</td>
</tr>
</tbody>
</table>

4. References
