Acoustic and perceptual analysis of vocal tremor

Christophe Mertens\textsuperscript{1}, Jean Schoentgen\textsuperscript{1,2}, Francis Grenez\textsuperscript{1}, Sabine Skodda\textsuperscript{3}

\textsuperscript{1}Laboratory of Images, Signals and Telecommunication Devices, Universite Libre de Bruxelles, Brussels, Belgium
\textsuperscript{2}National Fund for Scientific Research, Belgium
\textsuperscript{3}Department of Neurology, Knappschaftskrankenhaus, Ruhr University of Bochum, Germany

chmerten@ulb.ac.be, jschoent@ulb.ac.be, fgrenez@ulb.ac.be, sabine.skodda@kk-bochum.de

Abstract

The objective is to present a method to analyze vocal tremor in sustained speech sounds and compare the results with a perceptual scoring of the degree of tremor. The vocal cycle lengths are tracked by salience analysis and dynamic programming. The cycle length time series is then split into three components by empirical mode decomposition: jitter, tremor and trend. Tremor cues are obtained via non-uniform quantization of the spectrum of the tremor component of the cycle length time series. The results report tremor size, tremor frequency and bandwidth of a corpus of vowels sustained by Parkinson speakers as well as their correlations with the perceived degree of tremor assessed via pairwise comparison.

Index Terms: speech analysis, laryngeal assessment, vocal tremor

1. Introduction

The framework of the presentation is the assessment of disordered voices. The assessment of voice and laryngeal function is based on auditory ratings and acoustic analyses of speech sounds. Acoustic feature-based assessment methods are indeed popular because they are non-invasive and enable clinicians to monitor the voice of patients quantitatively.

Fast, small and involuntary cycle-to-cycle perturbations of vocal cycle lengths are designated as vocal jitter and involuntary low-frequency modulations of the vocal cycle lengths are referred to as vocal tremor. The latter have physiological (breathing, cardiac beat and pulsatile blood flow) or neurological causes. Conventionally, vocal jitter and tremor are tracked in sustained speech sounds in which small cycle length perturbations are less likely to be masked by intonation or accentuation. Disorders of phonation are often a consequence of the inability of vocal folds to vibrate normally. Larger than normal disturbances of the periodicity of the glottal source signal are therefore observed frequently as a consequence of organic or functional disorders of the larynx.

The objective of the study that is presented here is to measure vocal tremor frequency and size in normal speakers and patients suffering from neurological diseases. The analysis relies on the tracking of the vocal cycle lengths in sustained voiced speech sounds by means of a temporal method that is not based on strong assumptions with regard to the regularity of the speech cycles and their periodicity \cite{1}. The obtained cycle length time series is then decomposed further into a sum of oscillating components by empirical mode decomposition \cite{2,3}. According to their frequency content, these modes are then assigned to three categories: trend, vocal tremor and vocal jitter. The length time series components that are so obtained are then further analyzed with a view to obtaining the tremor size and frequency as well as jitter size.

Tremor and jitter size are correlated with scores assigned by three judges to the perceived degree of tremor. The relevance of numerical cues of vocal cycle lengths perturbations may indeed be evaluated by their ability to predict subjective scores that are obtained via the auditory assessment of the vocal timbre. In this study, a perceptual rating that is based on comparative judgments \cite{4} of the perceived degree of tremor in all possible signal pairs is used to rank according to perceived tremor level voiced speech samples sustained by 15 speakers.

The speakers have been Parkinson patients. Parkinson’s disease is a degenerative disorder of the central nervous system. During the initial stages of the disease, the symptoms are shaking, rigidity and slowness of limb motion. Possible vocal symptoms of the disease are vocal frequency tremor and hoarseness \cite{6}.

2. Method

2.1. Corpus

The corpus comprises sustained vowels [a] produced by 15 Parkinson speakers (3 female and 12 male speakers between 50 and 75 years of age). The signals have been recorded at a sampling frequency of 44.1 kHz in WAV format in the same recording environment and by means of the same equipment at the Department of Neurology of Bochum University Clinic. A stable speech signal fragment of 3 seconds has been selected from each vowel for analysis.

2.2. Perceptual assessment by pairwise comparison of tokens

In the framework of a rating session, all possible pairs of the 15 stimuli are presented randomly to a listener who is asked to designate the token of the pair with the highest perceived level of tremor. The listener can also designate the two tokens of a pair as equally perturbed. The total number of pairs is equal to 15 \times 14 \times 2. The software that presents the pairs one after the other increments by one the score of each token that is
designated by the listener as the most perturbed. The increment is equal to 0.5 when both tokens are declared having the same level of tremor. At the end of a listening session, the speech samples are ranked according to their total score. The advantage of scoring by pairwise comparison is that the overall ranking of the stimuli is obtained on the base of the ability of the listeners to compare two stimuli rather than on the base of their ability to categorize the stimuli according to a subjective scale. Three male listeners familiar with the analysis and scoring of disordered voices, but without any training in neurology have participated in the rating sessions. The three participants report normal hearing.

2.3. Tracking of the vocal cycle lengths

The vocal cycle length tracking is based on a temporal method, which does not rest on the assumptions that the signal is locally periodic and the average cycle length is known a priori. The vocal frequency is assumed to be comprised between 60 Hz and 400 Hz.

The cycle length tracking rests on the detection of cycle patterns that characterize the same glottal event. The selection of the length time series among several candidate patterns relies on dynamic programming that extracts a cycle sequence of length perturbations of which are minimal [1]. The cost function involves the second order differences of successive speech cycle durations as well as the cycle peak saliences. A peak is defined as a signal sample the amplitude of which is larger than its neighbors and the peak salience is defined as the length of the longest temporal interval over which a peak is a local maximum. The so obtained vocal cycle length time series is then constant-step resampled for further processing.

2.4. Categorization of vocal cycle lengths perturbations

The vocal cycle length time series is split into three components: trend (intonation, declination), vocal tremor and vocal jitter, with a view to analyzing vocal cycle length perturbations. A desirable property of empirical mode decomposition when compared to band-pass filtering is that the speech cycle time series can be reconstructed without losses by summing the empirical modes. Empirical mode decomposition consists in the break-up of the vocal cycle length time series \(x(n)\) into a sum of \(M\) functions (called intrinsic mode functions, IMF) and a monotonic function \(r(n)\) (called residue), as follows:

\[
x(n) = \sum_{i=1}^{M} c_i(n) + r(n)
\]  

An intrinsic mode function \(c_i(n)\) is an oscillating function with respect to the local average of the time series that is decomposed. IMFs must therefore satisfy two conditions:

1. The number of signal extrema and the number of zero crossings are equal or differ by one.
2. The average of the upper and lower intrinsic mode function envelope must be equal to zero at each discrete time instant \(n\).

Here, each mode function is assumed to be characterized at most by two dominant frequencies. These are found by low-pass filtering the cepstrum of the mode spectrum at 1/7.5 s and detecting all peak positions and heights in the so-obtained spectral contour \(C_i(f)\). The most prominent peak \((f_m,C_i(f_m))\) is also determined. Mode categorization is then carried out as follows.

1. The peaks which are located in the frequency interval \(\leq 20\) Hz and the amplitudes of which are \(\geq \frac{C_i(f_m)}{2}\) and the bandwidths of which that are \(\leq 6\) Hz are assigned to cycle length tremor.
2. The cycle length tremor time series is then obtained by means of the sum of all the mode functions \((c_i(n), i = k \ldots M)\) which have been assigned to the cycle length tremor category.

\[
x_{\text{tremor}}(n) = \sum_{i=k}^{M} c_i(n)
\]  

3. The cycle length jitter time series is then obtained on the base of the remaining mode functions.

\[
x_{\text{jitter}}(n) = \sum_{i=1}^{k-1} c_i(n)
\]  

4. The vocal trend, finally, is considered to be equal to the residue \(r(n)\).

\[
x_{\text{trend}}(n) = r(n)
\]

Figure 1 illustrates obtained slow and fast cycle length perturbations for a fragment of sustained vowel [a] pronounced by a Parkinson speaker.

2.5. Fundamental frequency \(F_0\)

The fundamental frequency \(F_0\) is computed via the inverse of the average of the vocal trend time series \(x_{\text{trend}}(n)\).

2.6. Perturbation levels

Vocal jitter and vocal tremor modulation depth (respectively noted \(\sigma_{\text{jitter}}\) and \(\sigma_{\text{tremor}}\) in \%) are computed via the standard deviation of their respective component time series \((x_{\text{jitter}}(n)\) or \(x_{\text{tremor}}(n)\)) divided by the average of the vocal trend time series \(x_{\text{trend}}(n)\). The total perturbation \((\sigma_{\text{pert}}, \%\) is obtained via the sum of the vocal jitter and vocal tremor time series.

2.7. Vocal tremor frequency

Perceptual evaluations show that at any time one prominent frequency at most is audible in the Parkinson speech stimuli. Hereafter, the most typical frequency is therefore detected in the tremor time series spectrum and used to characterize tremor. It is obtained via non-uniform quantization of the amplitude spectrum of the tremor component of the cycle length time series. In addition, two other cues are computed. They report the ratio of global tremor level explained by the typical tremor frequency and inform on the bandwidth of the latter.

2.7.1. Non-uniform quantization

The goal of the non-uniform quantization is the mapping of the frequency axis of the spectrum to a set of frequency intervals and their center of gravity, using Lloyd’s algorithm [5]. The steps are the following.

1. Initialization : \(j = 0\)
2. The frequency axis is split into a number $L$ of intervals $(I_{i,j})$, $i = 1 \ldots L$ that are regularly spaced in frequency between 0 and $15\,\text{Hz}$. The number of frequency intervals is related to the length $N$ of the Hanning window $w(n)$ used for the spectrum estimation (main lobe bandwidth of $[W(e^{j\omega})] = 8\pi/N$). Assuming that 3 frequency intervals are comprised in the window main lobe bandwidth, the number of intervals $L$ is fixed as follows.

$$ L = \left\lceil 3 \cdot \frac{2\pi \cdot 15 \, \text{Hz}}{8\pi/N} \right\rceil $$

3. The abscissa of the center of gravity $f_{G(i,j)}$ of each frequency interval $I_{(i,j)}$ is computed.

4. On the base of the positions of $f_{G(i,j)}$, $L$ new frequency intervals $I_{(i,j+1)}$ are fixed. The left and right boundaries $b_{(i,j+1)}$ ($i = 0 \ldots L$) of $I_{(i,j+1)}$ are given by the average of 2 successive positions of $f_{G(i,j)}$. The first and last boundaries ($b_{(0,j+1)}$ and $b_{(L,j+1)}$) are not fixed at $0\,\text{Hz}$ and $15\,\text{Hz}$ but are obtained by mirroring as follows to decrease boundary effects.

$$ \begin{align*}
  b_{(i,j+1)} &= \frac{f_{G(i,j)} + f_{G(i+1,j)}}{2}, & i &= 1 \ldots L - 1 \\
  b_{(0,j+1)} &= \max(0\,\text{Hz}, 2 \cdot f_{G(1,j)} - b_{(1,j+1)}) \\
  b_{(L,j+1)} &= \min(15\,\text{Hz}, 2 \cdot f_{G(L,j)} - b_{(L-1,j+1)})
\end{align*} $$

5. Steps 3 and 4 are repeated ($j \rightarrow j + 1$) till the interval boundaries stay fixed. The $L$ final frequency intervals obtained by non-uniform quantization are labelled as $I^*_i$ and the abscissa of their centres of gravity are labelled as $f_{G(i)}$.

2.7.2. Selection of a typical tremor frequency

The typical tremor frequency is determined on the base of the frequency intervals $I^*_i$, $i = 1 \ldots L$. The steps are the following:

1. Initialization: The mean energy $E_{(i)}$ of each frequency interval $I^*_i$, as well as the mean energy $E_T$ of the low-frequency spectrum ($[0, 15\,\text{Hz}]$) are computed.

2. The frequency intervals the mean energy of which are 10% larger than $E_T$ are selected and their center frequencies $f_G$ are considered to be vocal tremor frequency candidates and kept in memory.

3. If several adjacent frequency intervals $I^*_i$, $i = k \ldots l$ are selected, they are combined into one. The new interval is characterized by only one tremor frequency candidate equal to the weighted average of all combined frequencies $f_{G(i)}$. The weights are the respective average interval energies $E_{(i)}$.

4. The frequency interval the energy of which is the largest is kept and its frequency $f_G$ used to characterize tremor.

2.7.3. Relevance of the typical tremor frequency

The relevance of the so obtained typical tremor frequency can be assessed on the base of two additional cues. The first one is the ratio $\alpha$ between the energy of the interval of the typical tremor frequency divided by the energy of the whole spectrum of the vocal tremor time series. The ratio of the global tremor level explained by the typical tremor frequency interval is then obtained as follows:

Figure 1: Categorization: Vocal jitter, vocal tremor and vocal trend time series (in temporal domain) for fragment of vowel [a] sustained by a Parkinson speaker.

Figure 2 shows the amplitude spectrum of a vocal tremor time series as well as the mean energy and centre frequency of each frequency interval $I^*_i$.
σα = \sqrt{σ_{tre}^2 \cdot α} \tag{7}

The second cue corresponds to the width of the typical tremor frequency interval.

3. Results and conclusion

The vocal tremor depth in % has been correlated with perceptual scores that have been obtained via a pairwise auditory assessment of the vocal tremor level.

Table 1 reports the Pearson’s linear correlation coefficients between scores obtained via comparative judgements by 3 male listeners (J1, J2 and J3). One observes a good inter-judge agreement of listeners.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( J_1 )</th>
<th>( J_2 )</th>
<th>( J_3 )</th>
<th>( J_{aver} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J_1 )</td>
<td>X</td>
<td>0.84</td>
<td>0.89</td>
<td>0.95</td>
</tr>
<tr>
<td>( J_2 )</td>
<td>0.00</td>
<td>X</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td>( J_3 )</td>
<td>0.00</td>
<td>0.00</td>
<td>X</td>
<td>0.96</td>
</tr>
<tr>
<td>( J_{aver} )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 1: Inter-listener agreement of listeners. Pearson’s linear correlation coefficients (above the diagonal) and corresponding probability values (below the diagonal) between scores obtained via comparative judgements by 3 male listeners (J1, J2 and J3). \( J_{aver} \) designates the average scores of the three judges.

Pearson’s linear correlation coefficients between the degree of perceived vocal tremor level and the measured vocal cues are given in Table 2. The cues are the vocal jitter and vocal tremor modulation depths \( σ_{jit} \) and \( σ_{tre} \), the total perturbation level \( σ_{pert} \) and the ratio \( σ_α \) of global tremor level explained by the selected typical tremor frequency.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( \sigma_{tre} )</th>
<th>( \sigma_α )</th>
<th>( \sigma_{jit} )</th>
<th>( \sigma_{pert} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J_{aver} )</td>
<td>X</td>
<td>0.84</td>
<td>0.79</td>
<td>0.19</td>
</tr>
<tr>
<td>( σ_{tre} )</td>
<td>0.00</td>
<td>X</td>
<td>0.91</td>
<td>0.34</td>
</tr>
<tr>
<td>( σ_α )</td>
<td>0.00</td>
<td>0.00</td>
<td>X</td>
<td>0.22</td>
</tr>
<tr>
<td>( σ_{jit} )</td>
<td>0.49</td>
<td>0.20</td>
<td>0.42</td>
<td>X</td>
</tr>
<tr>
<td>( σ_{pert} )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 2: Pearson’s linear correlation coefficients (above the diagonal) and the corresponding probability values (below the diagonal) between the degree of perceived vocal tremor and the vocal perturbation cues. \( J_{aver} \) designates the average listener scores. \( σ_{tre}, σ_{jit} \) and \( σ_{pert} \) designate vocal tremor, vocal jitter and the total perturbation. \( σ_α \) refers to the ratio of the tremor level explained by the typical tremor interval.

One observes that the vocal tremor level \( σ_{tre} \) is significantly correlated \((r = 0.84)\) with the average listener scores. The high correlation coefficient between \( σ_{tre} \) and the total perturbation level \( σ_{pert} \) \((r = 0.97)\) suggests that the voice quality of Parkinson speakers is mainly affected by vocal tremor rather than hoarseness. The correlation between perceptual scores and the ratio of tremor level explained by the typical tremor frequency \( σ_α \) is high \((r = 0.79)\) but less than the correlation with \( σ_{tre} \). This is expected because other frequency components (i.e. physiological tremor) contribute to the total tremor energy.

One observes also a poor correlation \((r = 0.19)\) between the perceptual scores and vocal jitter level \( σ_{jit} \), which suggests that hoarseness and vocal tremor levels may be analysed and perceptually assessed separately.
4. References


