ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Analysis of vocal tremor and jitter by empirical mode decomposition of glottal cycle length time series

C. Mertens [1], *F. Grenez* [1], *J. Schoentgen* [1,2]

[1]Laboratory of Images, Signals and Telecommunication Devices,
Faculte des Sciences Appliquees, Universite Libre de Bruxelles, Brussels, Belgium
[2] National Fund for Scientific Research, Belgium

chmerten@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be

10.21437/Interspeech.2012-449

Abstract

The presentation concerns a method for tracking cycle lengths in voiced speech and breaking up vocal cycle length fluctuations in cycle length jitter and cycle length tremor. The tracking of the cycle lengths is based on a dynamic programming algorithm, which does not request that the signal is locally periodic and that the average period length is known a priori. The cycle length time series are decomposed into a sum of intrinsic mode functions by means of empirical mode decomposition. These mode functions are then assigned to three phenomena, which are cycle length jitter, cycle length tremor and trend owing to intonation and physiological tremor. We report tests of the proposed analysis by means of synthetic disordered speech sounds and illustrate slow and fast cycle length perturbations in modal and essential tremor speakers.

Index Terms: vocal frequency, vocal tremor, vocal jitter, speech salience analysis, empirical mode decomposition

1. Introduction

The objective is to present an analysis method of the cycle lengths in sustained speech sounds with a view to characterizing vocal jitter and vocal tremor.

The speech cycle tracking we implement does not rest on the assumptions that the speech signal is locally periodic and the average period length known a priori. We propose instead to track speech cycles via a multi-scale analysis that assigns a salience to each local speech signal peak. The salience of a local speech signal peak designates the time interval over which it is a maximum [2]. The vocal cycle detection relies on dynamic programming to extract a cycle length sequence the perturbations of which are minimal. The cost function involves the second order differences of successive speech cycle durations as well as the cycle peak saliences [3].

The cycle length time series is decomposed into a sum of empirical modes by means of empirical mode decomposition [1]. The high-frequency modes ($15Hz << F_0/2$) are assigned to vocal jitter, the low-frequency ones ($3Hz << 15Hz$) to vocal tremor and the ultra-low frequency modes ($< 3Hz$) are assigned to physiological tremor and declination.

2. Method

The analysis involves the following steps : preprocessing, sample salience analysis, speech cycle tracking, constant-step resampling of the cycle lengths, empirical mode decomposition, and tremor frequency and depth characterization via cepstral analysis of the tremor modes.

2.1. Preprocessing

The speech signal is band-pass filtered by means of a finite response (FIR) filter with cut-off frequencies equal to 60Hz and 1000Hz to remove additive low and high-frequency noise as well as high-frequency formants. The speech signal is upsampled to 192kHz to enable the speech cycle peak positions 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 method consists in assigning a salience value to each signal sample via a multi-scale analysis. The sample salience is defined as the length of the longest temporal interval over which a 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. For instance, in voiced speech, 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 multi-scale analysis is based on a sliding analysis window of length N. Fast windowed salience analysis involves speeding up the algorithm by computing left and right-hand saliences for a subset of samples only [2].

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, that 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. For speech cycle tracking, only the signal peak salience values are kept. A peak is any signal sample larger than its left and right neighbours.

2.3. Speech cycle tracking

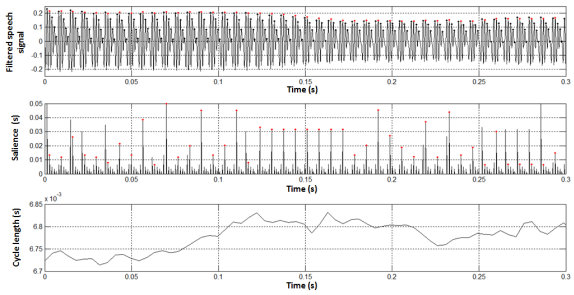

Figure 1: Example of cycle length tracking in voiced speech : Fragment of vowel [a] (above), signal peak saliences (middle), cycle length time series (bottom). Red markers indicates retained speech cycle peaks.

For speech cycle tracking, no strong assumptions are made with regard to the regularity of the cycle lengths. The vocal frequency is assumed to be comprised between $60Hz$ and $400Hz$. Vocal cycle detection relies on dynamic programming to extract a cycle sequence the length perturbations of which are minimal [3]. The cost function involves the second order differences of successive speech cycle durations as well as the cycle peak saliences.

The first stage consists in ranking the signal peaks according to decreasing salience and keeping those peaks the saliences 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 inter-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.

2.4. Constant-step resampling

The obtained vocal cycle length time series is then constant-step resampled by reconstructing the temporal axis as the sum of the successive vocal cycle lengths D_i (see Figure 2), interpolating the obtained series by means of cubic splines and resampling to obtain a time series of lengths sampled at a constant sampling frequency equal to 150% of the average vocal frequency.

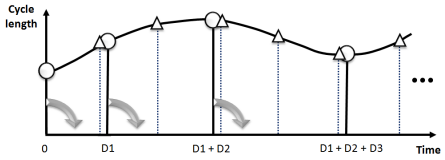

Figure 2: Constant-step resampling : reconstruction of time axis. Symbol $\circ$ reports the non-resampled time series and symbol $\triangle$ reports the constant-step resampled time series

2.5. Empirical mode decomposition

Empirical mode decomposition, developed by N.E. Huang in 1998, is a time-frequency analysis method. As opposed to Fourier decomposition that decomposes a signal into basis functions, empirical mode decomposition breaks up the signal $x(n)$ into a sum of M alternating functions $c_i(n)$ (called *intrinsic mode functions*, IMF) and a monotonic function $r(n)$ (called *residue*). An intrinsic mode function $c_i(n)$ is defined as an oscillating and symmetric function with respect to the local average of the time series that is analyzed. An intrinsic mode function must therefore satisfy the 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 envelope must be equal to zero at each instant n.

2.6. Vocal Jitter, vocal tremor and trend

On the base of the M extracted mode functions, three time series that are related to cycle length jitter, cycle length tremor and vocal trend are obtained. The categorization is carried out on the base of the spectrum of each mode function. The modes that form the cycle length tremor time series are selected as follows. Tremor frequencies are assumed to be in the interval $[3Hz, 15Hz]$ and the amplitude spectrum of each mode is obtained. For each mode spectrum the following tests are carried out.

1. Assuming that each mode function is characterized at most by two dominant frequencies, these frequencies are found by low-pass liftering the cep-

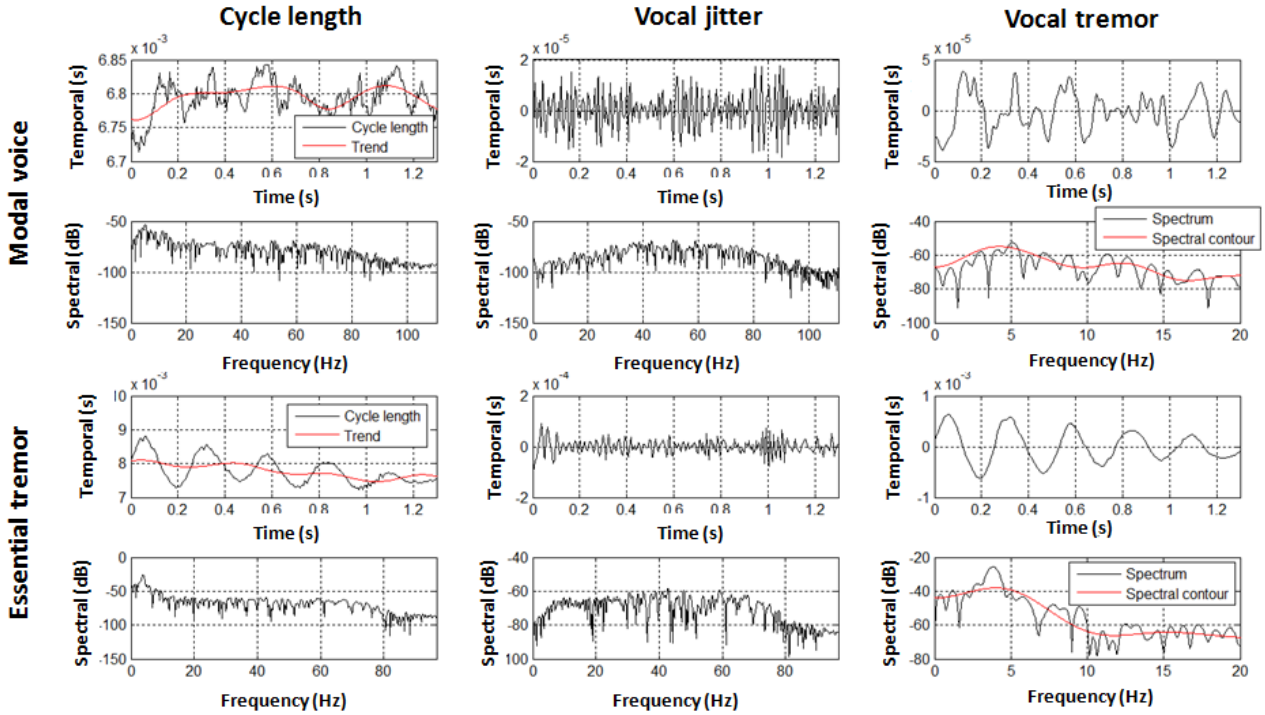

Figure 3: Categorization : Vocal jitter, vocal tremor and vocal trend time series (in temporal and spectral domain) for fragment of vowel [a] sustained by modal and essential tremor speakers

strum of the mode spectrum at $1/7.5$ s and finding all peak positions in the so-obtained spectral contour $C_i(f)$. The most prominent peak $(f_m, C_i(f_m))$ is also determined.

2. The peaks which are located in the frequency interval $[3Hz, 15Hz]$ and the amplitudes of which are $\geq \frac{C_i(f_m)}{\sqrt{2}}$ and the bandwidths $\leq 6Hz$ are selected.

3. The selected mode functions $c_i(n)$ are assigned to the cycle length tremor category and the positions of the retained peaks are kept in memory for further processing.

The cycle length tremor time series is given by the sum of all the mode functions $(c_i(n), i = k \ldots l)$ which have been assigned to the cycle length tremor category.

$$x_{tremor}(n) = \sum_{i=k}^{l} c_i(n) \qquad (1)$$

The cycle length jitter time series and the vocal trend (intonation and physiological tremor) time series are then obtained on the base of the remaining mode functions :

$$x_{jitter}(n) = \sum_{i=1}^{k-1} c_i(n) \qquad x_{trend}(n) = \sum_{i=l+1}^{M} c_i(n) + r(n) \qquad (2)$$

where $r(n)$ is the residue of the decomposition.

For instance, Figure 3 illustrates obtained slow and fast cycle length perturbations in modal and essential tremor speakers.

2.7. Vocal cues

The cycle length tremor frequencies are found by determining the frequencies of the peaks in the contour obtained by cepstral analysis of the spectrum of the cycle length tremor series. The cut-off quefrency is related to the number of different frequencies which have been obtained during categorization. Previously per mode discovered frequencies are considered to report distinct tremor frequencies when they differ by more than $3Hz$. This distance is selected on the base of the assumption that per mode at most two frequencies are dominant and that their bandwidths is $\leq 6Hz$. Indeed, one observes that distinct modes do not report always distinct frequency candidates. The final selection of the tremor frequency(ies) therefore rests on the contour of the amplitude spectrum of the tremor time series.

When more than one tremor frequency is discovered, their weighted average is obtained and reported as (a single) tremor frequency. The weights are the peak heights in the tremor spectral contour. The vocal tremor modulation depth and vocal jitter in % are computed via the stan-

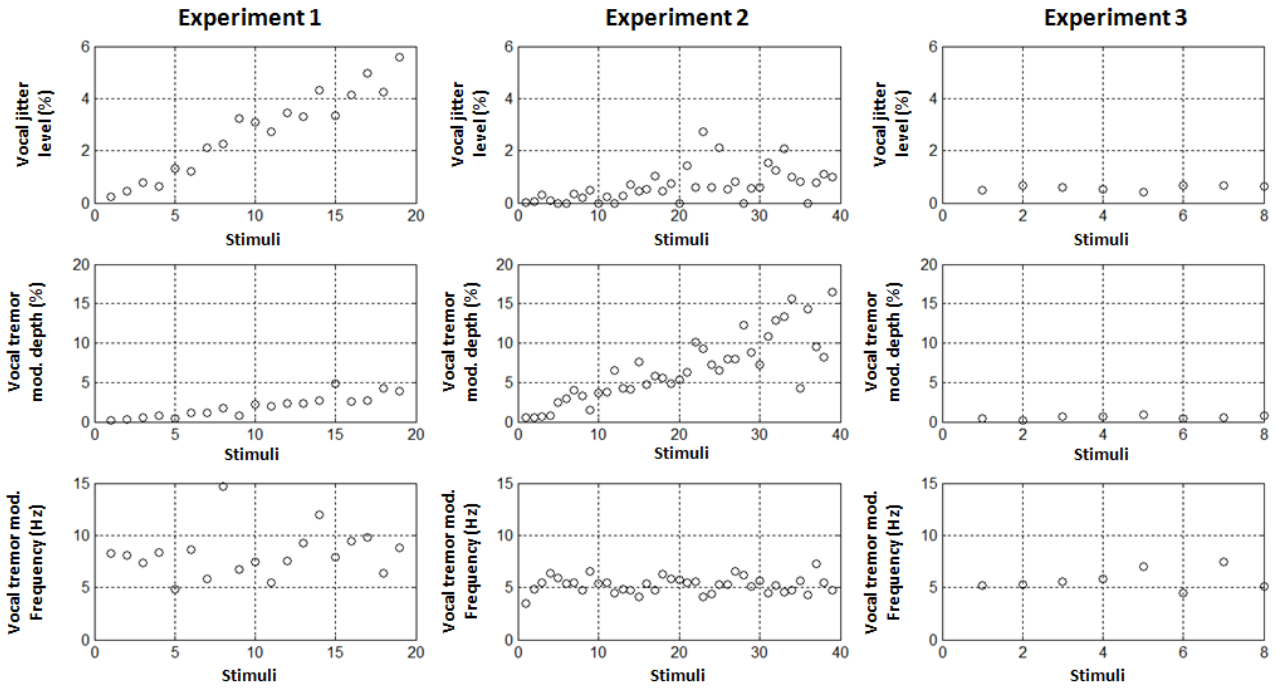

Figure 4: Evolution of vocal cues with different vocal jitter (left), vocal tremor (middle) and additive noise (right) characteristics. Horizontal axes report signal labels, vertical axes report jitter in %, frequency tremor depth in % and tremor frequency in Hz

dard deviation of the cycle length time series divided by the average of the empirical mode decomposition residue, which reports intonation.

2.8. Corpora

The method has been applied to several corpora of synthetic vowels [a]. Synthetic vowels [a] have been generated with increasing vocal jitter (19 stimuli), increasing vocal tremor (39 stimuli) and increasing additive noise (8 stimuli), with a view to validating the present approach. The average vocal frequency is fixed at 100Hz, the default vocal tremor frequency is fixed at 4Hz and the default vocal jitter is fixed at a very low level.

3. Results and discussion

Here, the reliability of the vocal cycle perturbation extraction has been tested by means of synthetic vowels, generated with different vocal jitter levels (Experiment 1), vocal tremor amplitudes (Experiment 2) and additive noise levels (Experiment 3). Figure 4 shows the results of these experiments. One observes, for experiments (1) and (2), that the standard deviation of the tremor time series increase with the modulation depth and the standard deviation of the jitter time series increase with vocal jitter. One observes also that the vocal tremor modulation frequency is correctly discovered as long as until the vocal jitter level is not much higher than the vocal tremor modulation depth (see Experiment 1).

4. References

[1] Norden E.Huang, Zheng Shen, Steven R.Long, Manli C.Wu, Hsing H.Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung and Henry H.Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proceeding of the The Royal Society, 1998.

[2] C.Mertens, F.Grenez and J.Schoentgen, Speech sample salience analysis for speech cycle detection, Proceedings 10th Annual Conference of the International Speech Communication Association INTERSPEECH, Brighton (U.K.), 2009

[3] C.Mertens, F.Grenez, L.Crevier-Buchman and J.Schoentgen, Reliable tracking based on speech sample salience of vocal cycle length perturbations, Proceedings 11th Annual Conference of the International Speech Communication Association INTERSPEECH, Makuhari (Japan), 2010

[4] Gabriel Rilling, Patrick Flandrin and Paulo Goncalves, On empirical mode decomposition and its algorithm, Proceedings of the 6th IEEE/EURASIP Workshop on Nonlinear Signal and Image Processing, Grado (Italy), 2003