Vocal tremor analysis via AM-FM decomposition of empirical modes of the glottal cycle length time series

Christophe Mertens, Francis Grenez, François Viallet, Alain Ghio, Sabine Skodda, Jean Schoentgen

Vocal tremor analysis via AM-FM decomposition of empirical modes of the glottal cycle length time series

Christophe Mertens[1], *Francis Grenez*[1], *François Viallet*[3,4], *Alain Ghio*[4], *Sabine Skodda*[5], *Jean Schoentgen*[1,2]

[1]Laboratory of Images, Signals and Acoustics. Université Libre de Bruxelles, Brussels, Belgium
[2]National Fund for Scientific Research, Belgium
[3]Neurology Department of Pays d'Aix Hospital, France
[4]LPL, Laboratoire Parole et Langage, CNRS UMR 6057, Aix-Marseille University, France
[5]Department of Neurology, Knappschaftskrankenhaus, Ruhr University of Bochum, Germany

chmerten@ulb.ac.be, fgrenez@ulb.ac.be, francois.viallet@lpl-aix.fr, alain.ghio@lpl-aix.fr, sabine.skodda@kk-bochum.de, jschoent@ulb.ac.be

Abstract

The presentation concerns a method that obtains the size and frequency of vocal tremor in speech sounds sustained by normal speakers and patients suffering from neurological disorders. The glottal cycle lengths are tracked in the temporal domain via salience analysis and dynamic programming. The cycle length time series is then decomposed into a sum of oscillating components by empirical mode decomposition the instantaneous envelopes and frequencies of which are obtained via an AM-FM decomposition. Based on their average instantaneous frequencies, the empirical modes are then assigned to four categories (intonation, physiological tremor, neurological tremor as well as jitter) and added within each. The within-category size of the cycle length perturbations is estimated via the standard deviation of the empirical mode sum divided by the average cycle length. The tremor frequency within the neurological tremor category is obtained via a weighted instantaneous average of the mode frequencies followed by a weighted temporal average. The method is applied to two corpora of vowels sustained by 123 and 74 control and 456 and 205 Parkinson speakers respectively.

Index Terms: Parkinson's disease, vocal tremor, empirical mode decomposition

1. Introduction

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, accentuation or segment-specific phenomena.

The objective of the presentation is to report estimates of the vocal tremor frequency and vocal tremor depth in vowels sustained by patients suffering from Parkinson disease and normal control speakers.

The analysis involves speech cycle features that are the cycle peak amplitudes and saliences, which generate candidate glottal cycle length time series. The final length time series is obtained via the selection by dynamic programming of a cycle length sequence the overall disturbance of which is a minimum.

The focus of the presentation is on the decomposition of the cycle length time series into sub time series that are respectively assigned to vocal jitter, neurological tremor, physiological tremor and a residual trend, which is due to intonation and declination, as well as the subsequent estimation of neurological tremor frequency and depth.

The break-up is based on empirical modes that are assigned to one of the four categories. Empirical modes are zero-mean alternating functions, which are obtained via empirical mode decomposition and the cycle length sub time series are obtained by summing the modes assigned to a same category.

The size of vocal jitter and of physiological and neurological tremor depths can so be estimated via the standard deviations of the corresponding sub time series.

Ideally, the typical tremor frequencies could be estimated via weighted averages of the instantaneous frequencies of the empirical modes that belong to the neurological or physiological tremor categories. However, one observes that due to mode mixing large frequency components may appear in two adjacent modes that together do not contribute much to the corresponding sub time series because they are out of phase. Computing arithmetic averages of mode frequencies would therefore assign undue weight to frequencies that exist in individual modes, but which do not contribute to the corresponding tremor category. The solution that is explored here consists in estimating the instantaneous phase and amplitude of each empirical mode. These then enable a complex mode to be assigned to each empirical mode and the weighted average of the instantaneous mode frequencies is defined in the complex plane so that only in-phase frequency components contribute to the final estimate.

Neurological tremor frequencies and depths are obtained for two corpora of vowels sustained by 123 and 74 control and 456 and 205 Parkinson speakers respectively. Parkinson's disease is a degenerative disorder of the central nervous system. Possible vocal symptoms of the disease are vocal frequency tremor and hoarseness [1]. In a study based on a large sample of patients with Parkinson disease, it has been reported that between 70% and 90% of patients have problems related to speech and voice impairments [2].

2. Methods

2.1. Cycle length tracking

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 known a priori. The vocal frequency is assumed to be between $60Hz$ and $400Hz$. The cycle length tracking relies on the amplitudes and saliences of the cycle peaks. Saliences are the lengths over which a cycle peak is a local maximum. The selection of the final length time series among several candidate series relies on dynamic programming that extracts a cycle sequence the length perturbations of which are minimal [3]. The so obtained cycle length time series is then constant-step resampled for further processing.

2.2. Empirical mode decomposition

The vocal cycle length time series is analyzed via Empirical Mode Decomposition. EMD breaks up iteratively a time series $x(t)$ into a sum of Intrinsic Mode Functions $c_i(t)$ and a residue $r(t)$.

$$x(t) = \sum_{i=1}^{I} c_i(t) + r(t) \tag{1}$$

An IMF is a zero-mean function alternating with respect to the horizontal axis and the residue is monotonic. The first IMF contains the finest scale components of the original time series and the successive IMFs contain increasing longer cycle variations. A property of empirical mode decomposition is that the original time series can be exactly recovered by summing the empirical modes. Another property is that the method does not rely on basis functions that are fixed a priori. The orthogonality between consecutive IMFs is not guaranteed theoretically, however. Therefore, mode mixing may occur, which means that 2 successive IMFs may overlap substantially in the frequency domain.

2.3. Instantaneous frequencies of IMFs and AM-FM decomposition

2.3.1. Overview

Given that IMFs are narrow-band functions locally, the instantaneous amplitude and frequency can be estimated for each. Each mode function $c_i(t)$ is therefore rewritten as a product.

$$c_i(t) = a_i(t)\ cos(\theta_i(t)) \tag{2}$$

The first factor, $a_i(t)$, is the time-varying IMF envelope (AM component or instantaneous amplitude) and the second, $cos(\theta_i(t))$, is the FM component, or carrier, with unit amplitude and instantaneous phase $\theta_i(t)$. A necessary condition for (2) to be meaningful is that $c_i(t)$ is mono-component and narrow-band so that the spectra of the AM and FM components do not overlap [4].

2.3.2. Empirical AM-FM decomposition

The empirical AM-FM decomposition is iterative and involves the following steps [4].

1. Initialization : $y(t) = c_i(t)$ and $a_i(t) = 1$, $\forall t$.
2. Detection of the local maxima of the value of $|y(t)|$.
3. Cubic spline interpolation of the envelope $e(t)$ of $|y(t)|$ via the positions and amplitudes of the local maxima.
4. Update : $a_i(t) \rightarrow a_i(t) \cdot e(t)$ and $y(t) \rightarrow \frac{y(t)}{e(t)}$.
5. Test whether all maxima of $|y(t)|$ have amplitude ≤ 1. If yes, stop. If no, loop to step 2.

Residual $y(t)$, which is the carrier, is requested to be ≤ 1 in absolute value because of identity (3). Result $a_i(t)$ corresponds to the time-varying envelope of $c_i(t)$ and the oscillating component $cos(\theta_i(t))$ is obtained via a division.

$$\cos\left(\theta_i(t)\right) = \frac{c_i(t)}{a_i(t)} \tag{3}$$

2.3.3. Computation of the instantaneous frequency

Instantaneous frequency $f_i(t)$ is obtained by a numerical differentiation of instantaneous phase $\theta_i(t)$ after phase unwrapping. The numerical phase differentiation relies on a 6^{th}-order polynomial [5].

$$f_i(t) = \frac{1}{2\pi}\ \frac{d\theta_i(t)}{dt} \tag{4}$$

2.4. Categorization of IMFs

The next step consists in grouping IMFs in four categories followed by summing : trend, physiological tremor, neurological tremor, and vocal jitter. Individual modes are assigned to one of the four categories on the base of their weighted instantaneous frequency averaged over the analysis interval. The weights are instantaneous amplitudes $a_i(t)$.

- The residue of the empirical mode decomposition is assigned to the trend, i.e. intonation and declination.
- The IMFs with weighted average instantaneous frequency $< 2Hz$ are assigned to physiological tremor.
- The IMFs with weighted average instantaneous frequency $\geq 2Hz$ and $\leq 15Hz$ are assigned to neurological tremor.
- The remaining IMFs are assigned to jitter.

$$x(t) = \underbrace{\left(\sum_{i=1}^{i_J} c_i(t)\right)}_{x_{jit}(t)} + \underbrace{\left(\sum_{i=i_J+1}^{i_N} c_i(t)\right)}_{x_{neur}(t)} + \underbrace{\left(\sum_{i=i_N+1}^{I} c_i(t)\right)}_{x_{physio}(t)} + \underbrace{r(t)}_{x_{trend}(t)} \tag{5}$$

Lower limit of 2Hz enables to discard heart beat and breathing, and the upper limit of 15Hz includes the frequency interval of the great majority of tremor types [6]. The categorization is illustrated in Figure 1.

2.5. Typical neurological tremor frequency

Here, only empirical modes assigned to the neurological tremor category are considered. The neurological tremor time series $x_{neur}(t)$ is given by the sum of the assigned modes $c_i(t)$.

$$x_{neur}(t) = \sum_{i=i_J+1}^{i_N} c_i(t) \tag{6}$$

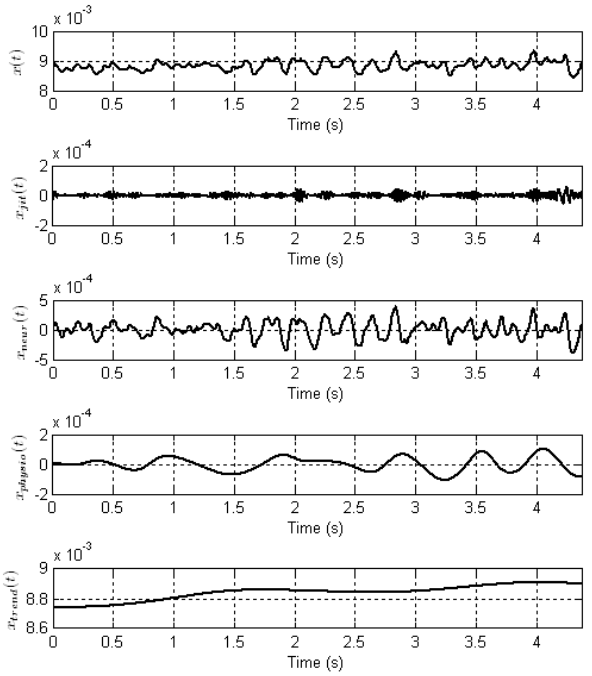

Figure 1: Categorization : cycle length time series, vocal jitter, neurological tremor, physiological tremor and trend (in that order) for a fragment of vowel [a] sustained by a Parkinson speaker.

In the complex domain, each time series $c_i(t)$ may be described on the basis of a complex time series $z_i(t)$ characterized by instantaneous phase $\phi_i(t)$ and envelope $a_i(t)$:

$$c_i(t) = a_i(t)\, cos\left(\phi_i(t)\right) = Re\left\{a_i(t)\, e^{j\phi_i(t)}\right\} = Re\left\{z_i(t)\right\} \tag{7}$$

As a consequence, a complex neurological mode $z_i(t)$ may be assigned to each empirical mode and a complex tremor time series $z_{neur}(t)$ is therefore assigned to real sum (6).

$$z_{neur}(t) = \sum_{i=i_J+1}^{i_N} z_i(t) = |z_{neur}(t)|\; e^{j\phi_{neur}(t)} \tag{8}$$

The derivative of complex time series (8) is calculated to obtain instantaneous frequency $\frac{1}{2\pi}\frac{d\phi_{neur}(t)}{dt}$. Assuming that the envelope evolves more slowly than the carrier, term $\frac{da_i(t)}{dt}$ may be disregarded, and, after some manipulations, one obtains:

$$f_{neur}(t) = \frac{1}{2\pi}\frac{d\phi_{neur}(t)}{dt} \approx \frac{\sum_{i=i_J+1}^{i_N} w_i(t)\, f_i(t)}{\sum_{i=i_J+1}^{i_N} w_i(t)} \tag{9}$$

with $w_i(t) = a_i(t)\, cos\big(\phi_i(t) - \phi_{neur}(t)\big)$

As a consequence, the typical neurological tremor frequency $f_{neur}(t)$ is given by the weighted sum of instantaneous empirical mode frequencies $f_i(t)$. The weights, $w_i(t)$ are obtained by projecting complex modes $z_i(t)$ on sum $z_{neur}(t)$. A large and positive weight is so assigned to empirical modes that are in phase with neurological tremor. Adjacent empirical modes that are out of phase (owing to mode mixing) zero each other and do not contribute to overall neurological tremor.

2.6. Vocal cues

2.6.1. Vocal frequency

Vocal frequency F_0 is obtained via the inverse of the average of the trend.

2.6.2. Perturbation levels

Vocal jitter σ_{jit}, neurological and physiological tremor modulation depths, σ_{neur} and σ_{physio}, total tremor depth σ_{tre} and total perturbation σ_{pert} are respectively the standard deviation of the jitter time series, the physiological and neurological tremor time series, the sum of the physiological and neurological tremor time series and the sum of the jitter and total tremor time series, divided by the average of the trend.

2.6.3. Typical neurological tremor frequency

Inspired by development (9), the typical neurological tremor frequency $\hat{f}_{neur}(t)$ is estimated on the base of a weighted sum of individual mode frequencies $f_i(t)$.

$$\hat{f}_{neur}(t) = \frac{\sum_{i=i_J+1}^{i_N} w_i(t)\, f_i(t)}{\sum_{i=i_J+1}^{i_N} w_i(t)} \tag{10}$$

Figure 2 illustrates a typical estimated neurological tremor frequency time series obtained for a stationary fragment of vowel [a]. One observes that the frequency $\hat{f}_{neur}(t)$ can be locally negative. One reason is the occasional negativity of weights $w_i(t)$, which is the consequence of particular phase relations between the complex sum of the modes and individual modes.

Another reason is the negativity of instantaneous mode frequencies $f_i(t)$. The instantaneous frequencies of individual modes are positive given their definition, but negativity may appear because of the iteration involved in the empirical AM-FM decomposition that may turn a mode inflection point into a pair of a local maximum and minimum. A local non-negative minimun or non-positive maximum causes the instantaneous frequency to become locally negative.

The estimated neurological tremor frequency time series is therefore smoothed by means of a moving average filter of length $0.2s$.

This instantaneous frequency time series is summarized by its temporal average $\mu_{\hat{f}_{neur}}$ and standard deviation $\sigma_{\hat{f}_{neur}}$ weighted sample-by-sample by module $|z_{neur}(t)|$.

3. Corpora

The corpora comprise vowels [a] sustained by normal speakers and patients suffering from Parkinson's disease. The first corpus [74 controls (42♂,32♀), 205 Parkinson speakers (129♂,76♀)]

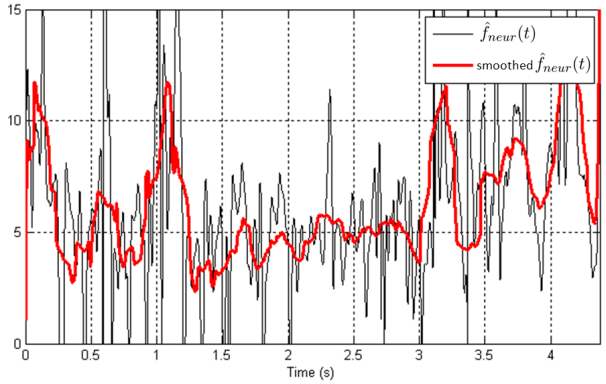

Figure 2: Raw and smoothed typical estimated neurological tremor frequency time series for a fragment of vowel [a] sustained by a Parkinson speaker

has 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. The second corpus [123 controls (50♂,73♀), 456 Parkinson speakers (302♂,154♀)] has been recorded at the Neurology Department of Pays d'Aix Hospital [7]. The samples of the Parkinson speakers group derive from patients in all the different stages of the disease (from mildly to very severely affected) and with very different disease durations.

The two corpora have been pooled with a view to analysing tremor depth and tremor frequency. The number of pooled control speakers has been 197 (92♂, 105♀) and the number of pooled Parkinson speakers has been 661 (431♂, 230♀).

4. Results and discussion

Table 1 reports the quartiles of the neurological and physiological tremor depths, the size of jitter, the size of the total tremor and of the total perturbations (all in %). The last three lines of the table report the tremor frequency, the tremor frequency standard deviation and the average F_0 (all in Hz). Quartiles Q_0 to Q_4 report the feature values at 5%, 25%, 50%, 75% and 95% of the feature value range of the pooled data.

A two-way analysis of variance has been carried out for each feature with independent variables "gender" and "pathology". Physiological tremor depth σ_{physio}, size of jitter σ_{jit} and neurological tremor frequency standard deviation $\sigma_{\hat{f}_{neur}}$ do not differ statistically significantly between control and Parkinson speakers or between male and female speakers. They are not discussed further.

The other features differed statistically significantly between control and Parkinson speakers. Neurological tremor depth σ_{neur} (< 0.05), neurological tremor frequency $\mu_{\hat{f}_{neur}}$ (< 0.001) and F_0 (< 0.001) also differed statistically significantly between male and female speakers. The interaction between variables "gender" and "pathology" was statistically significant for neurological tremor depth σ_{neur} (< 0.05) and F_0 (< 0.001).

Interactions between "gender" and "pathology" were due to the following. A non-parametric Wilcoxon test showed that F_0 differed statistically significantly between male control and Parkinson speakers only and that neurological tremor depth σ_{neur} differed between female control and Parkinson speakers only. In addition, F_0 increased for male Parkinson speakers and decreased for female speakers compared to the control speakers (loose from any statistical significance).

The remaining features, total tremor σ_{tre} (in %), total perturbations σ_{pert} (in %) and neurological tremor frequency $\mu_{\hat{f}_{neur}}$ (in Hz) increased for the Parkinson speakers compared to the control speakers, irrespective of the gender.

		MALE		FEMALE	
		CTRL	PARK	CTRL	PARK
$\sigma_{neur}(\%)$ $(\star\star\star)$	Q_0	0.48	0.5	0.33	0.41
	Q_1	0.76	0.77	0.59	0.76
	Q_2	1.03	1.02	0.79	1
	Q_3	1.37	1.44	1	1.44
	Q_4	2.76	2.92	1.88	3.02
$\sigma_{physio}(\%)$	Q_0	0.3	0.25	0.23	0.3
	Q_1	0.55	0.53	0.48	0.59
	Q_2	0.77	0.75	0.78	0.81
	Q_3	1.03	1.06	1.14	1.13
	Q_4	2.12	2.46	1.93	2.91
$\sigma_{jit}(\%)$	Q_0	0.14	0.14	0.13	0.14
	Q_1	0.23	0.29	0.24	0.28
	Q_2	0.35	0.4	0.35	0.41
	Q_3	0.57	0.76	0.55	0.65
	Q_4	2.11	2.88	2.85	2.35
$\sigma_{tre}(\%)$ $(\star\star)$	Q_0	0.63	0.67	0.46	0.64
	Q_1	0.98	0.99	0.84	1.02
	Q_2	1.32	1.34	1.17	1.34
	Q_3	1.73	1.8	1.53	1.88
	Q_4	3.51	3.53	2.68	3.69
$\sigma_{pert}(\%)$ $(\star\star)$	Q_0	0.66	0.72	0.61	0.67
	Q_1	1.05	1.09	0.93	1.13
	Q_2	1.4	1.49	1.29	1.46
	Q_3	1.89	1.99	1.66	2.07
	Q_4	3.86	4.29	3.27	4.23
$\mu_{\hat{f}_{neur}}(Hz)$ $(\star\star\star)$	Q_0	2.72	3.15	2.84	3.05
	Q_1	3.94	4.23	3.86	3.90
	Q_2	4.54	4.91	4.28	4.49
	Q_3	4.99	5.70	4.93	5.28
	Q_4	8.00	7.93	6.87	7.24
$\sigma_{\hat{f}_{neur}}(Hz)$	Q_0	0.83	1.15	1.24	1.20
	Q_1	1.69	1.80	1.77	1.68
	Q_2	2.15	2.25	2.04	2.14
	Q_3	2.60	2.65	2.47	2.48
	Q_4	3.63	3.63	3.58	3.74
$F_0(Hz)$ $(\star)$	Q_0	83	87	135	123
	Q_1	104	112	163	165
	Q_2	115	128	184	179
	Q_3	129	146	200	198
	Q_4	173	220	265	243

Table 1: Quartiles of the neurological and physiological tremor depth, the size of jitter, the size of the total tremor and of the total perturbations (all in %). The last three lines of the table report the tremor frequency, the tremor frequency standard deviation over the analysis interval and the average F_0 (all in Hz). Quartiles Q_0 to Q_4 report the feature values at 5%, 25%, 50%, 75% and 95% of the feature value range. Symbol $\star$ refers to the statistical significance of the differences between control and Parkinson speakers [$\star$: $p < 0.05$, $\star\star$: $p < 0.01$, $\star\star\star$: $p < 0.001$] within the framework of the two-way analysis of variance.

5. References

[1] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Lefebvre, and F. Grenez, "Low-frequency vocal modulations in vowels produced by parkinsonian subjects," *Speech Communication*, vol. 50, no. 4, pp. 288–300, 2008.

[2] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

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

[4] N. E. Huang, Z. Wu, S. R. Long, K. C. Arnold, X. Chew, and K. Blank, "On instantaneous frequency," *Advances in Adaptive Data Analysis*, vol. 01, no. 02, pp. 177–229, 2009.

[5] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. ii. algorithms and applications," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 540–568, 1992.

[6] L. Rubchinsky, A. Kuznetsov, V. Wheelock, and K. Sigvardt. (2015) Tremor. [Online]. Available: www.scholarpedia.org/article/Tremor

[7] A. Ghio, G. Pouchoulin, B. Teston, S. Pinto, C. Fredouille, C. D. Looze, D. Robert, F. Viallet, and A. Giovanni, "How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers?" *Speech Communication*, vol. 54, no. 5, pp. 664 – 679, 2012, advanced Voice Function Assessment.