A NOVEL METHOD FOR THE EXTRACTION OF VOCAL TREMOR

Yannis Pantazis, Maria Koutsogiannaki and Yannis Stylianou
Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
email: {pantazis, mkoutsog, yannis}@csd.uoc.gr

***Abstract*—Vocal tremor is defined as slow modulation of fundamental frequency or its amplitude [1], [2]. Even though vocal tremor may be attributed to neurological diseases, it may also be a natural stochastic modulation of voice. Many studies try to measure these modulations assuming that they are stationary. Hence, their analysis were limited to small intervals loosing important information about vocal tremor. We propose a novel method for the estimation of the modulations which is able to adapt to nonstationary environments. The method is mainly based on a AM-FM decomposition algorithm which is able to estimate the instantaneous components of speech signals. Results confirm that the method successfully extract the modulations of large speech segments and robustly estimate the time-varying modulation frequency and the time-varying modulation level.**

***Index Terms*—Voice quality, Vocal tremor, AM-FM decomposition**

I. Introduction

Typically, tremor in phonation is defined as modulations of the fundamental frequency and modulations of the amplitude due to the inability of humans to keep constant the tension of their vocal folds [3]. This phenomenon affects the glottal cycle in voiced speech making the fundamental frequency and the amplitude to vary stochastically. Vocal tremor is usually categorized into the physiological tremor which is a slow natural modulation of glottal cycle and the pathological tremor which is attributed to neurological diseases such as Parkinson or tremor of the limbs [4], [2]. Most importantly, while physiological tremor makes speech sound more natural and possibly more individual, pathological tremor may influence the quality of patients voice, hence, may influence the ability of patient's communication.

Moreover, while pathological tremor is characterized by stronger periodical patterns –a property that vibrato singing style has, too–, physiological tremor is more stochastic [4]. Even though, the analysis of physiological tremor is of great importance since vocal tremor in normophonic speakers may be an early sign of a neurological disease [5], [6]. Thus, it is useful to develop a toolkit that is able to measure or extract the vocal tremor even for normal voices. In the literature, acoustic analysis of tremor is usually based on the accurate estimation of fundamental frequency and then the characterization of the fundamental frequency's variations [1], [2]. Modulation frequency and modulation level are prominent attributes that are extracted from the instantaneous fundamental frequency [1], [2].

However, there are some issues not addressed in previous studies. Indeed, many studies are intersted only for the 1st harmonic which is related with the fundamental frequency but not for the higher harmonics. But 1st harmonic may be modulated by first formant which may lead to biased results. A more serious limitation of the previous studies is that the analyzed sustained vowel has duration that is one to two seconds. The reason for using short duration is that the modulation frequencies as well the modulation levels should be constant in order to apply classical frequency estimation analysis. This is a real drawback since the analysis of larger segments of speech may show interesting properties on vocal tremor [7], [8].

The objective of this article is to present and validate a novel method for the estimation of the vocal tremor on sustained vowels uttered by normophonic subjects. The proposed method assumes speech as a sum of time-varying sinusoids whose instantaneous amplitude and instantaneous frequency are estimated using a recently proposed AM-FM decomposition algorithm [9], [10]. The prime advantage of this algorithm is its ability to demodulate multicomponent signals (like speech) very accurately. Interestingly, any of the instantaneous components can be used for the analysis of vocal tremor and not only the 1st harmonic. Then, the second step of the algorithm is to subtract from the analyzed instantaneous component the very slow modulations ($< 2Hz$) in order to reveal the higher frequency modulations. This is achieved by filtering the instantaneous component using a Savinzky-Golay smoothing filter [11]. The final step is to estimate the modulation frequency and the modulation level which now are time-varying attributes because the modulations are primarily nonstationary. The estimation is performed using the same AM-FM decomposition algorithm applied for the extraction of instantaneous components. Results on sustained vowels uttered by normophonic speakers showed that the proposed method accurately estimate the instantaneous components of speech signals and then robustly extract the time-varying modulation frequency and modulation level.

The organization of the paper is as follows. Section II presents the tremor analysis method while in Section III the results are shown. Finally, Section IV concludes the paper.

II. Estimation of Vocal Tremor

Speech signals are modeled as a sum of time-varying sinusoidal components

$$s(t) = \sum_{k=1}^{K} a_k(t) cos(\phi_k(t)) \quad (1)$$

where K is the number of components, while $a_k(t)$ and $\phi_k(t)$ are the instantaneous amplitude and instantaneous phase of the k^{th} component, respectively. Moreover, instantaneous frequency, $f_k(t)$ is defined as the first derivative over time of the instantaneous phase:

$$f_k(t) = \frac{1}{2\pi}\frac{d\phi_k(t)}{dt} \quad (2)$$

In order to extract the characteristics of vocal tremor, the first crucial step is the estimation of the instantaneous components, $\{a_k(t), f_k(t)\}_{k=1}^{K}$ from the speech signal. The second step is to estimate and then remove the very slow modulation from the analyzed instantaneous component. The third and final step is the extraction of the modulation frequency and modulation level which are the vocal tremor characteristics we are interested.

A. Step 1: Estimation of Instantaneous Components

The estimation of the instantaneous components is achieved by an AM-FM decomposition algorithm recently proposed by Pantazis et al. [9], [10]. The AM-FM decomposition algorithm (AQHM in abbreviation) is an adaptive algorithm which is based on a time-varying sinusoidal model. The sinusoidal model is called quasi-harmonic model (QHM) and its parameters are estimated frame-by-frame through linear Least Square method. The prominent property of QHM is its ability to capture and then correct frequency estimation errors, thus, even if the analysis is performed in wrong frequencies, QHM is able to estimate the frequency mismatches and iteratively eliminate them.

The initialization of AQHM algorithm necessitates a rough estimate of the analysis frequencies for the first frame. During this study, the initial frequencies were assigned as integer multiples of an estimated fundamental frequency computed using the autocorrelation function of the first frame [12]. Moreover, time resolution of AQHM is determined by the hop-size of the algorithm while frequency resolution is determined by the window type and window length. We choose hop-size of $5ms$ and Hamming window as window function. The window's duration is adaptive and it was chosen to be three times the period of the smaller frequency.

The main advantage of AQHM algorithm is its ability to adapt to the signals characteristics. Indeed, after the first pass of the signal with QHM, an estimate for the instantaneous frequencies and instantaneous amplitudes, $\{\hat{a}_k(t), \hat{f}_k(t)\}_{k=1}^{K}$, have been obtained. Then, in the following passes, AQHM algorithm adapts the QHM basis functions using the estimated instantaneous frequencies. Thus, the bias due to the nonstationarity of the AM-FM signal is reduced and more accurate estimates for the instantaneous components are obtained. As an example, Fig. 1 shows the five first estimated instantaneous frequencies of a sustained vowel uttered by a male speaker using the AQHM algorithm.

B. Step 2: Removal of Very Slow Modulations

After choosing which instantaneous component will be analyzed, the second step of the analysis is to eliminate modulations which are less than $2Hz$. The removal of the trend is necessary in order to reveal the quasi-periodical modulations attributed to vocal tremor (compare Fig. 2a before elimination and Fig. 3a after elimination). However, before the removal of the trend, as a preprocessing step, we downsample the instantaneous component to have sampling frequency 1000 Hz. Indeed, since we are interested for modulations which are less than $20Hz$, the downsampled instantaneous component do not miss any important information.

The smoothing of the instantaneous component is performed using the Savinzky-Golay (S-G) filter [11], [13]. S-G smoothing filter essentially performs a local polynomial regression on a distribution of equally spaced points to determine the smoothed value for each point. The main advantage of this approach is that it tends to preserve features of the distribution such as relative maxima, minima and width, which are usually "flattened" by other adjacent averaging techniques like moving averages. The order of the local polynomial used in this study was 4 while the frame size was set $1s$ (1000 samples). Fig. 2a shows the instantaneous component as well its smoothed version for a sustained vowel. Fig. 2b implies that S-G filter captures the frequencies that are less than $2Hz$. Note that using different parameters for the S-G filter the smoothed fignal will capture more or less of the signal's frequencies. Then, the smoothed instantaneous component is subtracted from the unsmoothed in order to reveal the remaining modulations of the component.

C. Step 3: Extracting Vocal Tremor Characteristics

The final step is the modeling and estimation of the remaining modulations. As already stated, these modulations are nonstationary, hence, FFT-based approaches are not appropriate for this task. We suggest modelling the remaining nonstationary modulations as an amplitude modulated and frequency modulated signal. Mathematically, it is given by

$$x(t) = m(t)cos(\psi(t)) \quad (3)$$

where $x(t)$ are the remaining modulations of the instantaneous components, $m(t)$ is the instantaneous amplitude which with the appropriate scaling corresponds to the modulation level and $\psi(t)$ corresponds to the instantaneous phase. Once again, instantaneous frequency is given by $\zeta(t) = \frac{1}{2\pi}\frac{d\psi(t)}{dt}$ and corresponds to the modulation frequency.

AQHM algorithm is again applied for the estimation of the instantaneous components, $m(t)$ and $\zeta(t)$. The initial frequency of the first frame was computed by the largest peak of the FFT of the first frame while hop-size was set to $1ms$. Hamming window and its duration was determined using the same rule as in Step 1. Fig. 3a shows the reconstructed signal obtained from AQHM algorithm while Fig. 4 shows the estimated modulation frequency and estimated modulation level.

III. Results

In this section, the output of the proposed method for vocal tremor analysis is presented for normophonic speakers. The method is validated on a database of normal voices developed

in our recording lab. 11 male and 5 female healthy subjects whose age varies between 23 and 45 were participated. Sustained vowels $/a/$, $/e/$, $/i/$, $/o/$ and $/u/$ have been recorded at $48kHz$ and then downsampled at $16kHz$. The duration of sustained vowels varies from $2s$ to $8s$ depending primarily on the speaker.

Illustratively, Fig. 1 shows the first five harmonics extracted from sustained vowel $/a/$ using AQHM algorithm (Step 1). The signal which is reconstructed from the instantaneous components has signal-to-reconstruction error of about $32dB$ which proves that the analysis is very accurate. For the total database, the average signal-to-reconstruction error was more than $30dB$. Fig. 1 shows also that the modulations of higher harmonics are more evident which explains the use of modulation level which is relative to the mean value of the instantaneous component.

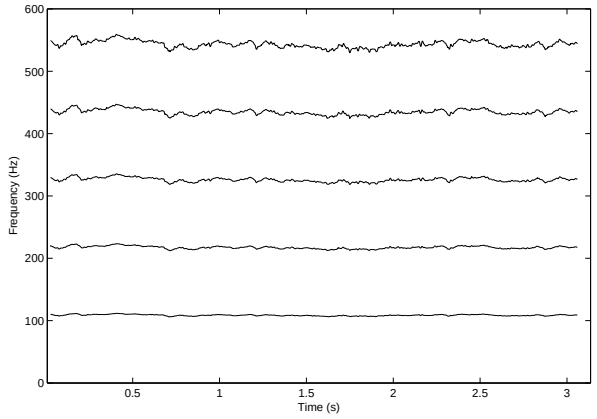

Fig. 1. First five instantaneous frequencies of a normophonic male speaker uttered the sustained vowel /a/.

Fig. 2 shows the instantaneous frequency of the 1st harmonic after removing its mean value and its filtered version using the S-G smoothing filter (Step 2). The smoothed instantaneous component contains information about the frequencies which are less than 2Hz. This component is then removed in order to reveal the modulations that are attributed to vocal tremor. Thus, the remaining component is analyzed using AQHM algorithm (Step 3). Fig. 3 indicates that the decomposition algorithm adapts to the nonstationary modulations of the signal. Extended tests on the database confirmed the ability of AQHM to adapt to the signal. The extracted time-varying modulation frequency as well the extracted time-varying modulation level are shown in Fig. 4. Modulation frequency takes values in this example between $2Hz$ and $13Hz$.

An important feature of the proposed method is that any of the instantaneous components can be analyzed. Fig. 5 and Fig. 6 shows the estimated modulation frequency and modulation level for the instantaneous amplitude of the 4th harmonic of the same sustained vowel. Interestingly, both modulation frequency and modulation level differs from them of Fig. 4.

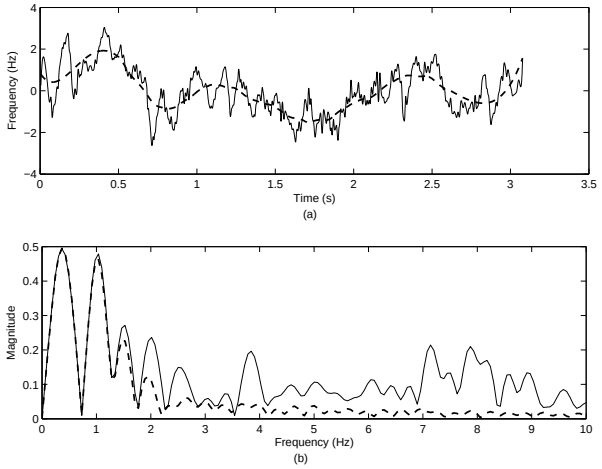

Fig. 2. (a) First harmonic of Fig. 1 without its mean value (continuous line) and its smoothed version (dashed line) are shown. (b) Fourier transform of signals in (a). S-G smoothing filter captures the frequencies that are below $2Hz$.

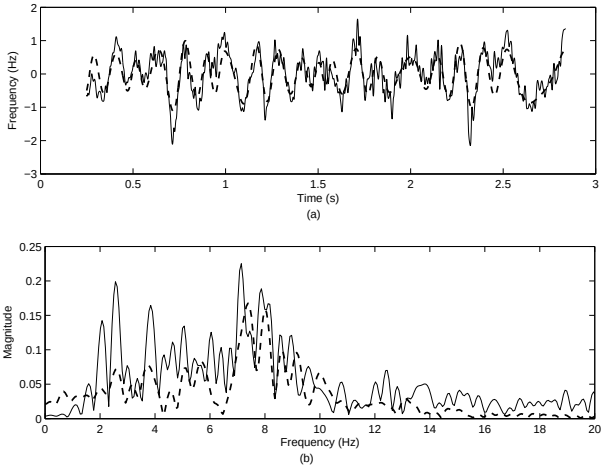

Fig. 3. (a) Instantaneous component after subtracting its smoothed version (continuous line) and the reconstruction of the AM-FM decomposition algorithm (dashed line). (b) Fourier transforms of the components in (a).

IV. Conclusion & Future Work

A novel method for the acoustical analysis of vocal tremor was presented. It is based on a AM-FM demodulation algorithm which is used for the extraction of both instantaneous amplitudes and instantaneous frequencies from the speech signal (Step 1) and the extraction of modulation frequency and modulation level from the analyzed instantaneous component (Step 3). Results indicate that the proposed method is cabable of handling large segments of sustained vowels where the assumption of modulations' stationarity is invalid and provide robust time-varying estimates for the modulation frequency and the modulation level.

Finally, while the proposed method was validated only on normophonic subjects it is of great importance to apply and test it in pathological subjects. Future work will be devoted

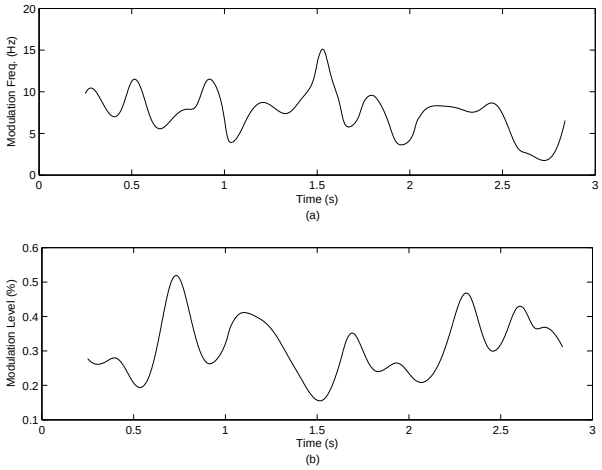

Fig. 4. (a) Modulation frequency of the signal in Fig. 3. (b) Modulation level of the same signal. Note that neither modulation frequency nor modulation level have constant values during the phonation.

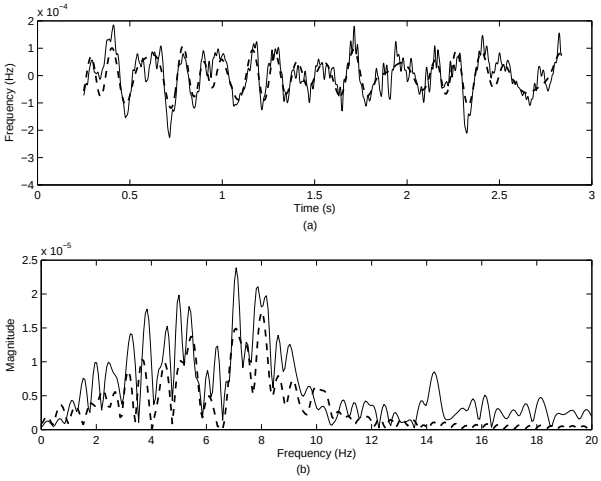

Fig. 5. Similar to Fig. 3 but for the instantaneous amplitude of the 4th harmonic. Note that the proposed vocal tremor extraction algorithm can be applied to any of the instantaneous component.

on analyzing pathological vocal tremor and possibly on other applications such as the analysis of vibrato singing style where the objective is to achieve a particular amount of modulation frequency and/or modulation level.

REFERENCES

[1] W. Winholtz and L. Ramig. Vocal Tremor Analysis with the Vocal Demodulator. 35:562–573, 1992.

[2] J. Schoentgen. Modulation Frequency and Modulation Level owing to Vocal Microtremor. *J. Acoust. Soc. Am.*, pages 690–700, Aug 2002.

[3] I. R. Titze. Motor and Sensory Components of a Feedback Control Model of Fundamental Frequency. *Producing Speech: Contemporaty Issues*, pages 309–320, 1995.

[4] H.J. Freund. *Central Rhythmicities in Moter Control and its Pertubances*, pages 79–82. Springer, Berlin, 1987.

[5] C.A. Meeuwis and E.A. Baarsma. Essential (Voice) Tremor. *Clinical Otolaryngology*, 5, 1985.

[6] L.J. Findley and M.A. Gresty. *Head, Face and Voice Tremor*, pages 239–253. New York: Raven, 1988.

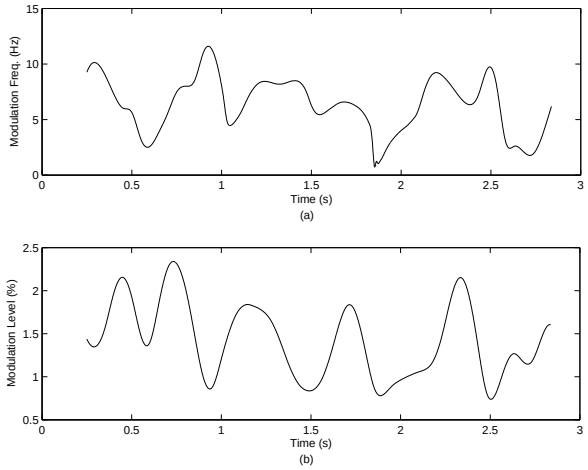

Fig. 6. Similar to Fig. 4 but for the instantaneous component of Fig. 5. Similarities and differences can be found between the modulation frequency and modulation level of instantaneous components.

[7] J. Kreiman, B. Gabelman, and B.R. Gerratt. Perception of Vocal Tremor. *Journal of Speech, Language and Hearing Research*, 46:203–214, 2003.

[8] H. Ackermann and W. Zeigler. Acoustic Analysis of Vocal Instability in Cerebellar Dysfunctions. *Annals of Otology, Rhinology and Laryngology*, 103:98–104, 1994.

[9] Y. Pantazis, O. Rosec, and Y. Stylianou. AM-FM Estimation for Speech based on a Time-varying Sinusoidal Model. In *Interspeech*, Brighton, 2009.

[10] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AM-FM Signal Decomposition with Application to Speech Analysis. *IEEE Trans. on Audio, Speech and Language Processing*, submitted.

[11] A. Savitzky and M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[12] T. F. Quatieri. *Speech Signal Processing*. Prentice Hall, Signal Processing Series, 2002.

[13] J. Steinier, Y. Termonia, and J. Deltour. Comments on smoothing and differentiation of data by simplified least square procedure. *Analytical Chemistry*, 44:1906–1909, 1972.