INTERSPEECH 2009 BRIGHTON

AM-FM Estimation for Speech Based on a Time-Varying Sinusoidal Model

Yannis Pantazis[1] , Olivier Rosec[2] and Yannis Stylianou[1]

[1]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
[2]Orange Labs TECH/SSTP/VMI, Lannion, France
pantazis@csd.uoc.gr, olivier.rosec@orange-ftgroup.com and yannis@csd.uoc.gr

10.21437/Interspeech.2009-24

Abstract

In this paper we present a method based on a time-varying sinusoidal model for a robust and accurate estimation of amplitude and frequency modulations (AM-FM) in speech. The suggested approach has two main steps. First, speech is modeled as a sinusoidal model with time-varying amplitudes. Specifically, the model makes use of a first order time polynomial with complex coefficients for capturing instantaneous amplitude and frequency (phase) components. Next, the model parameters are updated by using the previously estimated instantaneous phase information. Thus, an iterative scheme for AM-FM decomposition of speech is suggested which was validated on synthetic AM-FM signals and tested on reconstruction of voiced speech signals where the signal-to-error reconstruction ratio (SERR) was used as measure. Compared to the standard sinusoidal representation, the suggested approach found to improve the corresponding SERR by 47%, resulting in over 30 dB of SERR.

Index Terms: Sinusoidal modeling, AM-FM demodulation, Speech analysis, Speech reconstruction

1. Introduction

Amplitude Modulated and Frequency Modulated (AM-FM) signals are present everywhere in natural sounds including the human voice. Focusing in speech, amplitude and frequency modulations are strongly related with the speech production mechanism; from the glottis to vocal tract and to the lips. Not all the frequency bands of speech are affected by these modulations in the same way neither these modulations are similar to all speakers.

In the literature, there are two major categories of algorithms in analyzing an AM-FM signal and estimating its time-varying components. The first category is through non-parametric time-frequency representations like Spectrogram or STFT, Wigner-Ville distribution [1] and Choi-Williams distribution [1] [2]. In the second category, a model for the analyzed signal is implicitly or explicitly assumed and then the proposed techniques are trying to estimate the instantaneous components of the signal (the terms 'demodulation', 'separation' and 'decomposition' are also used in the same context). Analytic signal estimated through Hilbert transform is used to demodulate an AM-FM signal [3] while Energy Operators like Teager-Kaiser energy operator [4] were applied for AM-FM demodulation [5] [6]. All the above ways manifest their limitations especially if the input signal, like speech, is a multicomponent signal [7]. Parametric approaches such as Fan-Chirp transform was recently suggested in [8] [9] for speech analysis trying to address the time-varying and multicomponent character of speech. However, important parameters such as the chirp rate, that controls the time-varying frequency characteristics of the signal, is usually computed separately from the other parameters of a model, which may lead to suboptimal solutions and inaccurate estimations.

In this paper, we suggest an iterative approach for estimating the instantaneous components of a time varying multicomponent AM-FM signal based on a time-varying sinusoidal model. The initial (or basic) model was introduced by Laroche et al. [10] for audio and speech analysis [11]. In [12], the model was re-introduced revealing its main properties. It was shown that it is equivalent to a time-varying quasi-harmonic representation of speech. In the following, the basic model will be referred to as Quasi-Harmonic Model, QHM. In this paper we extend the use of QHM in decomposing a speech signal into its instantaneous amplitude and frequency components by suggesting an iterative scheme. More specifically, at each iteration we suggest to update the basis functions where the input signal is projected by using the instantaneous phase information estimated in the previous iteration. This will be referred to as iQHM. The suggested approach for AM-FM decomposition was validated on synthetic multi component AM-FM signals and tested on voiced speech signals. Examples of speech signal reconstruction are provided, showing the accuracy in estimating the AM and FM components in speech.

The organization of the paper is as follows. Section 2 overviews the main properties of the basic QHM model and provides the motivation of using the model for the decomposition of a signal into AM and FM components. In Section 3 the computation of the instantaneous components is developed and the iterative QHM (iQHM) is described. Experiments on synthetic multi-component AM-FM signals and of high-quality voice speech signal reconstruction are presented in Section 4. Section 5 concludes the paper.

2. AM-FM signals and QHM

Let us consider a time varying multi-component AM-FM signal

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

where $K(t)$ is the number of components, and $a_k(t)$ and $\phi_k(t)$ are the instantaneous amplitude (AM) and phase modulation (PM) of the k^{th} component, respectively. Another important time-varying component of $s(t)$ is its instantaneous frequency (FM), $\nu_k(t)$, which is defined as the first derivative over time of the instantaneous phase

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

Let us further assume that the analysis of the AM-FM signal is performed frame-by-frame, and that frame l occurs at time t_l:

$$s_l(t) = s(t - t_l)w(t) \quad (3)$$

where $w(t)$ is the analysis window which is zero outside a symmetric interval $[t_l - t_0, t_l + t_0]$ (t_l is considered as the center of analysis window). We suggest to model the l^{th} frame of $s(t)$, as

$$y(t) = \left(\sum_{k=-K}^{K} (a_k + tb_k) e^{2\pi j f_k t} \right) w(t) \quad -t_0 \le t \le t_0, \tag{4}$$

where we shifted the analysis interval to the origin, i.e., $t = t - t_l$ while we considered the number of components, K, to be constant over the analysis window. In (4), a_k and b_k denote the complex amplitude and complex slope of the k^{th} component. Assuming that K and frequencies f_k are known, the estimation of a_k and b_k is obtained by minimizing the mean-squared error between the model and the input signal. This leads to a simple least squares solution [11].

To better understand the connection between AM-FM signals and QHM we suggest further discussing the time and frequency domain properties of QHM.

2.1. Time-Domain Properties of QHM

From (4), it is easily seen that for each component the instantaneous amplitude for frame l is a time-varying function and it is given by

$$\begin{aligned} M_k(t) &= |a_k + tb_k| \\ &= \sqrt{(a_k^R + tb_k^R)^2 + (a_k^I + tb_k^I)^2} \end{aligned} \tag{5}$$

where x^R and x^I mean the real and imaginary parts of x, respectively. The instantaneous phase is given for each component by

$$\begin{aligned} \Phi_k(t) &= 2\pi f_k t + \angle(a_k + tb_k) \\ &= 2\pi f_k t + atan \frac{a_k^I + tb_k^I}{a_k^R + tb_k^R} \end{aligned} \tag{6}$$

while the instantaneous frequency is given by

$$\begin{aligned} F_k(t) &= \frac{1}{2\pi} \frac{d\Phi_k(t)}{dt} \\ &= f_k + \frac{1}{2\pi} \frac{a_k^R b_k^I - a_k^I b_k^R}{M_k^2(t)} \end{aligned} \tag{7}$$

2.2. Frequency-Domain Properties of QHM

QHM can be written in frequency domain as:

$$Y(f) = \sum_{k=-K}^{K} \left(a_k W(f - f_k) + j b_k W'(f - f_k) \right) \tag{8}$$

where $W(f)$ is the Fourier transform of the analysis window, $w(t)$ and $W'(f)$ is the derivative of $W(f)$ over f. In [12], the projection of b_k to a_k

$$b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k \tag{9}$$

was suggested where ja_k denotes the perpendicular (vector) to a_k. Also it was shown that for small values of $\rho_{2,k}$ the kth component of $Y(f)$ can be approximated by:

$$Y_k(f) \approx a_k \left[W(f - f_k - \frac{\rho_{2,k}}{2\pi}) + j\rho_{1,k} W'(f - f_k) \right] \tag{10}$$

Going back to the time domain, the kth component of $y(t)$ is written as:

$$y_k(t) \approx a_k \left[e^{j(2\pi f_k + \rho_{2,k})t} + \rho_{1,k} t e^{j2\pi f_k t} \right] w(t) \tag{11}$$

In (11), the term $\rho_{1,k}$ admits the following expression:

$$\rho_{1,k} = \frac{\frac{dM_k(t)}{dt}|_{t=0}}{M_k(0)} \tag{12}$$

and thus provides the normalized slope of the instantaneous amplitude of the kth component at the center of the analysis window. The term $\rho_{2,k}$ can be expressed as

$$\rho_{2,k} = \frac{a_k^R b_k^I - a_k^I b_k^R}{M_k^2(0)}$$

and then from (7) it follows that:

$$F_k(0) = f_k + \frac{\rho_{2,k}}{2\pi} \tag{13}$$

Thus, $\rho_{2,k}$ is the mismatch between the analysis frequency f_k and the value of the instantaneous frequency at the center of analysis window. In [12], $\rho_{2,k}$ was used to adjust the analysis frequencies, thus providing robust and accurate representation of nearly-harmonic signals like speech [12].

2.3. Motivation

The motivation of using QHM for AM-FM decomposition stems from the Taylor series expansion of the instantaneous phase of kth component in (1). At the center of the analysis window (i.e. relative time equals to zero) we have:

$$\phi_k(t) = 2\pi\zeta_k t + \sum_{i=0}^{\infty} \phi_{k,i} \frac{t^i}{i!} \tag{14}$$

where ζ_k is the carrier frequency while $\phi_{k,i}$ are the Taylor series coefficients. $\phi_{k,0}$ accounts for the phase offset of the kth component while $\phi_{k,1}$ can be viewed as a frequency mismatch of the carrier frequency. Then, from (2) and (14), the instantaneous frequency of the kth component is given at the center of the analysis window by

$$\nu_k(0) = \zeta_k + \frac{\phi_{k,1}}{2\pi} \tag{15}$$

Let us assume that the carrier frequencies, ζ_k, are known or have been estimated and set $f_k = \zeta_k$. We suggest then that the frequency mismatch, $\phi_{k,1}$, can be estimated by the correction term $\rho_{2,k}$ of QHM: $\hat{\phi}_{k,1} = \rho_{2,k}$. If our suggestion is valid, this leads to a simple method for the estimation of the instantaneous frequency $\nu_k(t)$ at the center of the analysis window (i.e., $\nu_k(0)$).

3. Computing the instantaneous components

We are now able to analytically develop the AM-FM decomposition algorithm. Let $a_{k,l}$ and $b_{k,l}$ denote the complex amplitude and slope estimated at time t_l. Then, given an initial estimate of the instantaneous frequency, $\hat{\nu}_k(t_0)$, the algorithm has two main steps.

For $l = 1, 2, ...$

1. Compute the $a_{k,l}$ and $b_{k,l}$ through Least Squares (LS) using $f_k^l = \hat{\nu}_k(t_{l-1})$.

2. Compute instantaneous components.
   For $k = 1 \ldots K$:
$$\begin{cases} \hat{a}_k(t_l) = M_k(0) & \text{for } t = 0 \text{ in (5)} \\ \hat{\phi}_k(t_l) = \Phi_k(0) & \text{for } t = 0 \text{ in (6)} \\ \hat{\nu}_k(t_l) = F_k(0) & \text{for } t = 0 \text{ in (7)} \end{cases}$$
3. Move to the next time instant t_{l+1} and go to 1).

The algorithm is intuitively simple, and, as concerns its complexity, the most time-consuming part is the computation of $a_{k,l}$ and $b_{k,l}$ via LS at each time step. The cost is $O((2K)^3 + 2KN)$, where N is the length of the window in samples and K is the number of components. For comparison purposes, when there is only one component, the complexity of each step is $O(2N)$ with N quite small, which is comparable to algorithms with very low complexity such as the DESA algorithm [5].

3.1. Importance of Window Length

From (4) and since f_k are considered to be known and a_k and b_k are complex numbers, it follows that there are 4K unknowns to estimate at each analysis frame, where K is the number of components. Then, the length of the analysis window in samples should be at least $4K$. Moreover, low frequency components need larger windows, and an empirical choice for the analysis window length is $2\lfloor \frac{f_s}{\min f_k} \rfloor$. Furthermore, when the AM-FM signal is contaminated by noise, more samples (larger window) are needed in order to perform more robust estimation. On the other hand, the approximation in (10) is less valid when larger windows are used. For that reason, we suggest as general rule to use window lengths as small as possible.

3.2. Iterative Estimation

Motivated by the results presented previously, we suggest an iterative algorithm for estimating the instantaneous amplitude and frequency components of a signal. First, we assume that the input signal was already analyzed by QHM and an estimate of the instantaneous phase $\hat{\phi}_k(t_l)$ for each component was obtained, where t_l denotes the centers of the analysis windows. From these phase estimations, we can furthermore estimate the evolution of phase information $\forall t$ by using interpolator functions like spline. However, to reduce the errors caused by such an interpolation approach, we perform analysis at every sample. We then suggest to update the analysis basis functions of QHM using the estimated instantaneous phase information, $\hat{\phi}_k(t)$:

$$y(t) = \left(\sum_{k=1}^{K} (a_k + t b_k) e^{j\hat{\phi}_k(t)} \right) w(t),\ t \in [-t_0, t_0] \quad (16)$$

The estimation of the complex amplitudes and complex slopes is again performed by Least Squares.

4. Validation and Testing

To support our suggestions, let us first consider a two-component AM-FM signal which contains sinusoidally time-varying amplitude and frequency components

$$\begin{aligned} s(t) = {} & 2(1 + 0.4cos(2\pi 30t))e^{j(2\pi 700t + cos(2\pi 130t))} \\ & + 2(1 + 0.3cos(2\pi 50t))e^{j(2\pi 1000t + cos(2\pi 130t))} \end{aligned} \quad (17)$$

which was constructed using a sampling frequency of $8000Hz$. It is worth noting that the modulation of the frequency component is considered rather fast (13 cycles in $0.1s$ period).

The instantaneous amplitude and frequency components of the input signal are depicted in Fig. 1 as solid lines. Applying the QHM analysis procedure on the input signal, a first estimation of these instantaneous components is obtained. For this, a squared Hamming window of duration $16ms$ was used and a hop size of one sample was applied. We assumed that there was a frequency mismatch between the input frequencies (ζ_k) and the analysis frequencies (f_k), of $32Hz$ for both components.

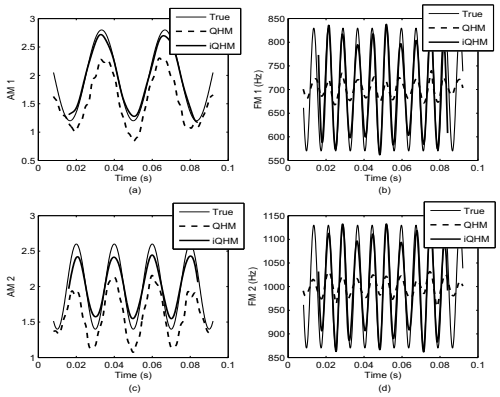

Figure 1: *The estimated and the true amplitude and frequency components of an AM-FM signal with sinusoidal amplitude and phase modulation.*

The estimated instantaneous amplitude and frequency components by QHM are also shown in Fig. 1 using bold dashed lines. Although the estimated amplitude components are comparable to the corresponding true information, the errors in the estimation of the instantaneous frequency components are not negligible. Nevertheless, we used this information in updating the bases function where the input signal is projected, as it was suggested in the iQHM algorithm. At every iteration, the instantaneous phase estimated in the previous iteration is used. After 15 iterations, the estimation of the instantaneous amplitudes and frequencies of the signal is highly improved as can be seen in Fig. 1 (bold solid line). More iterations were not found to further improve the results. Performance details for QHM and iQHM, using the mean absolute error in the estimation of the amplitudes and the frequencies of the input signal, are provided in Table 1.

Table 1: *Mean Absolute Error for QHM and iQHM for the two-component signal in (17).*

	AM1	AM2	FM1 (Hz)	FM2 (Hz)
QHM, 16ms	0.36	0.37	69.99	69.74
iQHM 15, 16ms	0.05	0.11	21.43	20.14

Let us now present results from speech analysis and reconstruction using the estimations of instantaneous amplitudes and frequencies provided by iQHM. Since we perform an analysis at every sample (hop size) of the signal, it would be interesting to compare the suggested approach to a much simpler and standard approach. For example, in each analysis frame, we may compute the Fourier transform of the windowed signal, and then by peak picking in the magnitude spectrum, to compute the amplitudes, frequencies and phases. The analysis window is then shifted by one sample and the above computations are

repeated. This is very similar to the standard analysis stage of the Sinusoidal Model (SM) [13]. Furthermore, the accuracy of peak picking approach is further improved by parabolic interpolation.We will refer to this approach as SM. Using the instantaneous amplitude and phase components, $\hat{a}_k(t)$, and $\hat{\phi}_k(t)$, respectively, the signal reconstruction is then simply provided as

$$s(t) = \sum_{k=1}^{K} \hat{a}_k(t) \cos\left(\hat{\phi}_k(t)\right) \tag{18}$$

In Fig. 2(a), a segment from a voiced speech signal generated by a male speaker is shown (sampling frequency 16kHz). All analyses were performed with a squared Hamming window of 16 ms. For QHM analysis, we set $f_k = kf_0$ with $f_0 = 140Hz$ (the average fundamental frequency of the segment) and $K = 40$. For SM analysis, 2048 frequency bins were computed and the most prominent 40 components were selected after peak picking and parabolic interpolation. We verified that the frequency of the selected peaks were closely related to the analysis frequencies, f_k of QHM. Regarding iQHM, only one iteration was used for this example. The estimated instantaneous amplitude and phase information for all the methods (SM, QHM, and iQHM) were then used to reconstruct the speech signal as in (18). The reconstruction error for each method is depicted in Fig. 2. As expected, iQHM provides the best reconstruction compared to the other two alternatives even if only one iteration is applied. The Signal-to-Error Reconstruction Ratio (SERR), is $19.5dB$ for SM, $24.1dB$ for QHM, and $30.5dB$ for iQHM.

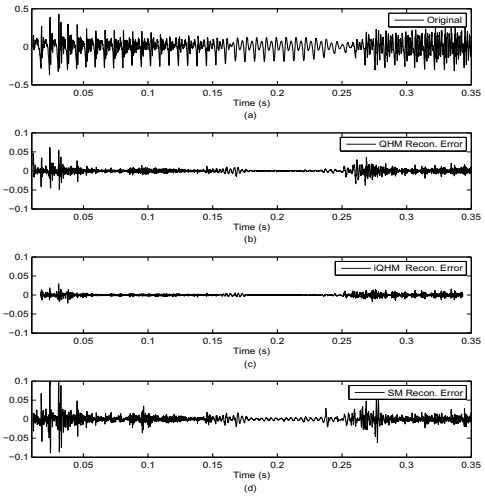

Figure 2: *(a) Original speech signal and reconstruction error for (b) QHM, (c) iQHM 3, and (d) SM*

We analyzed approximately one minute of voiced speech per gender (male and female speakers), using the same procedures as for the example shown above. The average score of each approach in terms of SERR (dB) is provided in Table. 2 along with the number of iterations used in iQHM. We observed that after 3 iterations the algorithm reached a stability in the reconstruction error power. From Table. 2, we observe that the reconstruction error has lower power in case of female voices than for male voices. This is expected as the distance between the components is larger in this case. Comparing iQHM after 3 iterations with SM, we see that the improvement in SERR is between 40% and 53%, thus providing an average improvement of 47%. Compared to QHM, the suggested approach provided an average improvement of 33% in SERR.

Table 2: *Signal-to-Error Ratio (in dB) for the proposed methods and sinusoidal model.*

	Male	Female
QHM	20.1	25.7
iQHM 1	26.0	31.5
iQHM 2	27.3	32.3
iQHM 3	27.8	32.8
SM	18.2	23.4

5. Conclusions

We presented a new method for the iterative estimation of AM-FM components of speech signals based on a time-varying sinusoidal model, referred to as QHM. The new method is referred to as iQHM. Signal re-synthesis using the instantaneous components leads to high-quality reconstruction of the signal with a Signal-to Reconstruction Error Ratio over 30dB showing the accuracy of the suggested estimator. Based on the obtained results, the proposed method is expected to be useful in many speech applications including speech analysis, speech synthesis and modifications and objective voice function assessment.

6. References

[1] L. Cohen. *Time-Frequency Analysis*. Prentice Hall, 1995.

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

[3] D. Vakman. On the Analytic Signal, the Teager-Kaiser Energy Algorithm, and other methods for defining Amplitude and Frequency. *IEEE Trans. on Signal Processing*, 44:791–797, 1996.

[4] J. F. Kaiser. On a Simple Algorithm to Calculate the 'Energy' of a Signal. In *Proc. IEEE ICASSP*, pages 381–384, Albuquerque, USA, Apr 1990.

[5] P. Maragos, J. Kaiser, and T. Quatieri. On Separating Amplitude from Frequency Modulations using Energy Operators. In *Proc. IEEE ICASSP*, pages 1–4, San Francisco, USA, Mar 1992.

[6] T. F. Quatieri T. E. Hanna and G. C. O'Leary. AM-FM Separation using Auditory-Motivated Filters. *IEEE Trans. on Speech and Audio Processing*, 5:465–480, 1997.

[7] J. K. Gupta, S.C. Sekhar, and T. V. Sreenivas. Performance Analysis of AM–FM Estimators. In *TENCON 2003*, pages 954–958, Oct 2003.

[8] L. Weruaga and M. Kepesi. The Fan-chirp Transform for Non-stationary Harmonic Signals. *IEEE Trans. on Signal Processing*, 87:1504–1522, 2007.

[9] R. Dunn and T. F. Quatieri. Sinewave Analysis/Synthesis based on the Fan-Chirp Transform. In *Proc. WASPAA*, pages 16–19, Oct 2007.

[10] J. Laroche. A new Analysis/Synthesis System of Musical Signals using Prony's Method. Application to Heavily Damped Percussive Sounds. In *Proc. IEEE ICASSP*, pages 2053–2056, Glasgow, UK, May 1989.

[11] J. Laroche, Y. Stylianou, and E. Moulines. HNM: A Simple, Efficient Harmonic plus Noise Model for Speech. In *Proc. WASPAA*, pages 169–172, New Paltz, NY, USA, Oct 1993.

[12] Y. Pantazis, O. Rosec, and Y. Stylianou. On the Properties of a Time-Varying Quasi-Harmonic Model of Speech. In *Proc. Interspeech*, pages 1044–1047, Brisbane, Australia, Sep 2008.

[13] R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34:744–754, 1986.