Time-Variant Least Squares Harmonic Modeling

Qin Li and *Les Atlas*
Department of Electrical Engineering, University of Washington
Seattle, WA 98195-2500, USA

ABSTRACT

An algorithm for harmonic decomposition of time-variant signals is derived from a least squares harmonic (LSH) technique. The estimates of harmonic amplitudes and phases are formulated as the solution of a set of linear equations which minimizing mean square error; the signal frequency is modeled by a linear or quadratic polynomial and obtained via a local search over polynomial coefficients. An initial estimate of signal frequency is necessary to reduce computation time. This method is capable of producing accurate and robust harmonic estimation in low SNR situations. We show applicability to high accuracy speech pitch and heart sound beat epoch estimation.

1. INTRODUCTION

Harmonic modeling, which is also known as a sinusoidal representation, has been widely applied in a number of areas such as speech coding, compression, enhancement, synthesis, and pitch estimation. Basically, harmonic modeling can be described as a finite combination of sinusoidal components, and was first introduced by McAulay and Quatieri [4].

Usually speech signals can be decomposed into two parts: a quasi-periodic (harmonic component) and a non-periodic part (noise component). A crucial step for harmonic modeling is to find the parameters of harmonic components, e.g. their amplitudes, frequencies, and phases. A variety of techniques have been proposed for this decomposition of harmonic signals, such as the high-resolution analysis of the short-time Fourier Transform (STFT) [4], the multiband excitation model (MBE) [3], and the least squares harmonic model (LSH) [1]. The first two techniques have been successfully used for low bit-rate speech coding; however their performance degrades at low SNR. The LSH model is capable of producing more accurate and robust harmonic analysis, even at very low SNR; however, as will be shown, its performance degrades significantly with rapid changes in signal frequency.

In this paper, we propose a harmonic analysis method for time-variant signals, which is a substantially modified version of the LSH approach developed by Abu-Shikhah and Deriche [1]. The key difference from LSH [1] is that the fundamental frequency of signal is allowed to vary with time within the data segment. A linear or quadratic polynomial is used to fit the frequency variation in the data segment. The best-fit harmonic estimation is then obtained via minimizing mean square error (MSE). As we will show, our extended LSH (ELSH) model has been successfully used to detect heartbeats from acoustic signals recorded from body-worn sensors with presence of very strong body-motion interference and ambient noise [7]. We also apply this technique to estimate pitch epochs from linear prediction (LP) residuals of speech signals. The quantitative modeling accuracy of harmonic estimation is also available as a reliable indicator to classify voiced versus unvoiced sounds.

2. EXTENDED LEAST SQUARES HARMONIC MODEL

Suppose we have a harmonic signal that consists of a set of sinusoids, which is given by

$$s(k)=\sum_{i=1}^{M} C_i \cos(i\omega_0(k)k+\phi_i)\,, \tag{1}$$

where $s(k)$ is a segment of a harmonic signal with length N. k=0,1, … , N-1 is the discrete time index, M is the total number of harmonic components, C_i and ϕ_i are the amplitude and phase angle for each harmonic component i=1,2, … , M, and $\omega_0(k)$ is the normalized fundamental frequency. For ELSH, our key difference from LSH [1], where ω_0 was constant, is that $\omega_0(k)$ is allowed to vary with time.

The LSH model, with our time-varying extension, assumes that the signal is the sum of a harmonic signal and a noise, so that the signal is given by

$$s(k)=h(k)+n(k)=\sum_{i=1}^{M} C_i \cos(i\omega_0(k)k+\phi_i)+n(k)\,. \tag{2}$$

The harmonic component $h(k)$ in equations (2) can be rewritten as:

$$\begin{aligned} h(k) &= \sum_{i=1}^{M} C_i \cos(i\omega_0(k)k+\phi_i) \\ &= \sum_{i=1}^{M} C_i\left[\cos(i\omega_0(k)k)\cos(\phi_i)-\sin(i\omega_0(k)k)\sin(\phi_i)\right] \\ &= \sum_{i=1}^{M}\left[A_i \cos(i\omega_0(k)k)-B_i \sin(i\omega_0(k)k)\right], \end{aligned} \tag{3}$$

where $A_i=C_i\cos(\phi_i)$, $B_i=C_i\sin(\phi_i)$, and

$$\phi_i=\begin{cases}\tan^{-1}\left(B_i/A_i\right), & \text{if } A_i\ge 0\\ \tan^{-1}\left(B_i/A_i\right)+\pi, & \text{if } A_i<0\end{cases}. \tag{4}$$

The weighted mean square error (MSE) between $s(k)$ and $h(k)$ is then:

$$\begin{aligned} E &= \frac{1}{N}\sum_{k=0}^{N-1}\{s(k)-h(k)\}^2 W(k) \\ &= \frac{1}{N}\sum_{k=0}^{N-1}\left\{s(k)-\sum_{i=1}^{M}\left[A_i \cos(i\omega_0(k)k)-B_i \sin(i\omega_0(k)k)\right]\right\}^2 W(k)\,, \end{aligned} \tag{5}$$

where $W(k)$ is weight of the k^{th} sample point for MSE calculation.

For a given sequence of fundamental frequency $\omega_0(k)$, the minimum MSE is found by

$$\frac{\partial E}{\partial A_j}=0\ ,\ \frac{\partial E}{\partial B_j}=0\ ,\ \text{for } j=1, 2, \dots, M. \tag{6}$$

The above equation ends up with a linear equation

$$\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix}=\begin{pmatrix}\mathbf{Q} & \mathbf{R}\\ \mathbf{S} & \mathbf{T}\end{pmatrix}\bullet\begin{pmatrix}\mathbf{A}\\ \mathbf{B}\end{pmatrix}\ , \tag{7}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $M\times 1$ unknown vectors to be determined, and other matrices are defined as:

$$\begin{aligned}
\mathbf{Q}(j,i)&=\sum_{k=0}^{N-1}\cos(i\omega_0 k)W(k)\cos(j\omega_0 k)\ ,\\
\mathbf{R}(j,i)&=-\sum_{k=0}^{N-1}\sin(i\omega_0 k)W(k)\cos(j\omega_0 k)\ ,\\
\mathbf{S}(j,i)&=\sum_{k=0}^{N-1}\cos(i\omega_0 k)W(k)\sin(j\omega_0 k)\ ,\\
\mathbf{T}(j,i)&=-\sum_{k=0}^{N-1}\sin(i\omega_0 k)W(k)\sin(j\omega_0 k)\ ,\\
\mathbf{Y}_1(j)&=\sum_{k=0}^{N-1}s(k)W(k)\cos(j\omega_0 k)\ ,\\
\mathbf{Y}_2(j)&=\sum_{k=0}^{N-1}s(k)W(k)\sin(j\omega_0 k)\ ,
\end{aligned} \tag{8}$$

where $i, j = 1, 2, \dots, M$, and $k=0, 1, \dots N-1$.

Solving equation set (7) for a given $\omega_0(k)$ results in

$$\begin{pmatrix}\mathbf{A}\\ \mathbf{B}\end{pmatrix}=\begin{pmatrix}\mathbf{Q} & \mathbf{R}\\ \mathbf{S} & \mathbf{T}\end{pmatrix}^{-1}\bullet\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix}\ , \tag{9}$$

and then C_i and ϕ_i can be obtained from equation (4).

Solving equation (7) only yields amplitudes and phases given a specified fundamental frequency $\omega_0(k)$. There is no closed-form analytical solution for the unknown input signal frequency. In order to estimate the unknown signal frequency, we need to repetitively solve equation (7) for a range of discrete $\omega_0(k)$'s, and select the $\omega_0(k)$ that gives the minimum MSE with corresponding C_i and ϕ_i as the final results. Since the result of a Fourier decomposition is unique, the true signal fundamental frequency $\omega_0(k)$ always yields the minimum MSE among all possible $\omega_0(k)$'s.

3. FUNDAMENTAL FREQUENCY ESTIMATE

If we allow $\omega_0(k)$ to vary independently for each $k=0,1,\dots,N-1$ and then search for all possible combinations of $\omega_0(k)$, the number of computations would be impractical. So here we choose a polynomial to model $\omega_0(k)$'s evolution in k (discrete time). The polynomial could be zeroth order (constant frequency, one free parameter), first order (linear chirp, two free parameters), second order (quadratic frequency, three parameters), or higher order. Since the computation time increases exponentially with polynomial order, the computation time for high order polynomials is considerable. An accurate initial estimation of frequency is necessary to narrower down the search range and lower computation time.

In order to lower computation time, we implemented the polymonial models via a step-up recursion. We start with a constant frequency model and search over all candidates of a range of discrete ω_0. This step is same as the algorithm describe in [1]. Then we step up with a linear frequency model. We formulate the instantaneous frequency of the signal as

$$\omega_{inst}(k)=a+bk\quad . \tag{10}$$

Since the instantaneous frequency is the derivative of the whole phase term $(i\omega_0(k)k+\phi_i)$ with respect to time, $\omega_0(k)$ in equation (2) is therefore written as

$$\omega_0(k)=a+\frac{1}{2}bk\quad . \tag{11}$$

For convenience, we set

$$k=\begin{cases}-\frac{N}{2}, \dots, -1, 0, 1, \dots, \frac{N}{2}-1 & \text{if } N \text{ is even}\\ -\frac{N-1}{2}, \dots, -1, 0, 1, \dots, \frac{N-1}{2} & \text{if } N \text{ is odd}\end{cases} \tag{12}$$

without losing loss of generalization, so that the instantaneous frequency $\omega_{inst}(k)$ has mean of a and slope of b. We use ω_0 calculated from the constant frequency model as the initial estimate of a. Then we search all combinations of a and b to find the one with minimum MSE.

If necessary, we can continue to step up to a quadratic time-varying frequency model, where we model instantaneous frequency as

$$\omega_{inst}(k)=a+bk+c(k^2-\frac{1}{N}\sum\nolimits_k k^2)\ , \tag{13}$$

and the corresponding phase term $\omega_0(k)$ is written as

$$\omega_0(k)=a+\frac{1}{2}bk+c(\frac{1}{3}k^2-\frac{1}{N}\sum\nolimits_k k^2)\ . \tag{14}$$

Adding a constant term $-\frac{1}{N}\sum_k k^2$ conveniently makes a the mean and b the slope of the instantaneous frequency. Therefore we can directly use a and b calculated from the linear model as the initial estimate.

We can ostensibly continue to step up to obtain high order frequency model via a similar formulation. However higher order frequency models require substantially more computation time and tend to fit noise. Using overlapped data windows also helps to obtain a better fit to frequency variation. In practice, the polynomial order should be determined from data quality, the nature of the unknown signal, and computation efficiency.

4. APPLICATIONS AND PERFORMANCE

Our extended LSH (ELSH) approach has been successfully applied to detect acoustic heartbeat signal at very low SNR. We also demonstrate its application to pitch estimation.

4.1 Acoustic Heartbeat Detection

For healthy and safety reasons there are needs for, say, an army or fire department to monitor soldier or firefighter's physiological indicators via body-worn acoustic sensors while they are doing their missions [7]. Heart rate is the most important physiological indicator for human health. Reliable algorithms are needed to estimate heart rate from acoustic heartbeat sounds. In such situations, the acoustic heartbeat signals: (1) have strong periodicity, but usually buried into

strong ambient noise and body motion interference; (2) are typically corrupted by additive noise sources that are non-stationary and diverse in structure; (3) have a rate which, due to varied activity, can change rapidly over a short time period. In other words, we are dealing with a harmonic analysis problem with potentially abrupt and rapid fundamental frequency change and low SNR.

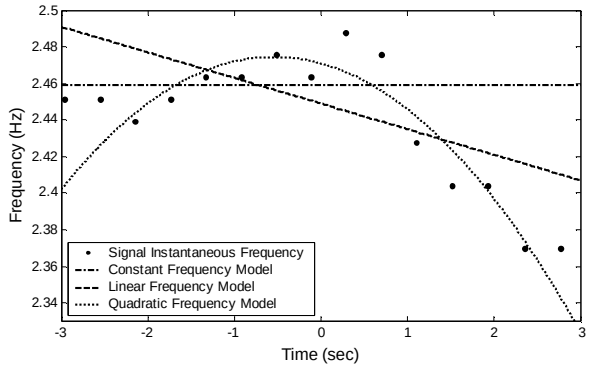

Figure 1. Comparison of modeling changes of instantaneous frequencies.

Frequency Model	Constant	Linear	Quadratic
Normalized Waveform MSE	54.0%	53.4%	12.0%

Table 1. Comparison of modeling errors

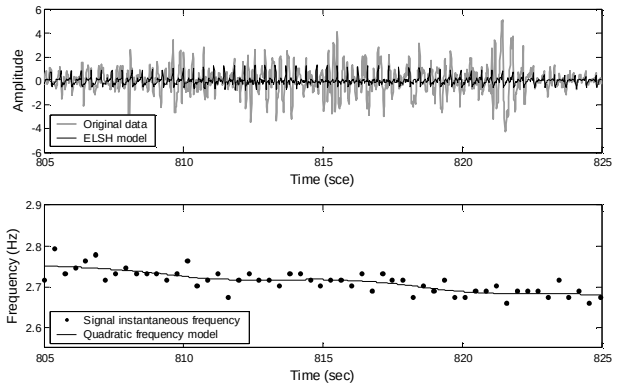

Figure 2. Results of ELSH approach on low SNR (-10 dB) data.

We first give an example for ELSH applied to a clean heartbeat signal (SNR = 10 dB) with a rapid heart rate change over a short time period. Figure 1 depicts the modeled instantaneous frequencies for the constant, linear and quadratic model, respectively. Table 1 shows the corresponding normalized mean square error (MSE) between the ELSH model and true waveform. From the constant to linear frequency model, there is a minor improvement; yet from linear to quadratic frequency model, the improvement is significant.

Another example is given for presence of colored noise with low SNR of -10 dB. In figure 2, the upper panel depicts the original signal and the ELSH model; the lower panel depicts the instantaneous frequency of the signal and our results from quadratic frequency model. The results clearly show that our harmonic analysis algorithm is very stable and reliable at low SNR.

4.2 Pitch estimation and voicing detection

In recent years, harmonic analysis methods have received much attention on speech coding [3, 6] and pitch estimation [2, 5]. We proposed a method to estimate the pitch frequency of speech signals using ELSH. The schematic diagram for pitch estimation is shown in figure 3. First a linear prediction (LP) analysis was performed for every 100 ms and a LP residual are generated. We partitioned the residual signal into 50 ms segments with a 25 ms overlap between segments to avoid discontinuities. The ELSH was then performed for each residual segment to obtain harmonic parameters and harmonic model resynthesis. A cos^2 windowing function was applied for accurate model reconstruction throughout overlapped regions of the segments.

Usually for speech coding or pitch estimation, a voiced-unvoiced (U/V) classifier is also needed. One of the advantages of our approach is that the U/V discrimination indicator can directly be obtained from harmonic modeling accuracy, e.g. a harmonic-to-noise ratio (HNR)

$$HNR = 10\log_{10}\left(\frac{\sum_k h(k)^2}{\sum_k [s(k)-h(k)]^2}\right), \quad (15)$$

which is very similar to HNR defined in [2]. Our HNR was calculated over a much shorter window (10 ms) to quickly follow transitions between voiced and unvoiced sounds.

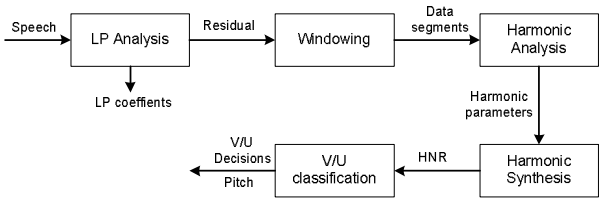

Figure 3. Schematic diagram of pitch estimation approach.

We tested the performance and robustness of the ELSH approach with polynomial time-varying frequency models. Speech test data was obtained from the Carnegie Mellon University speech group website, http://www.festvox.org/dbs/dbs_time.html. We hand marked pitch epochs for a comparison reference. ELSH was performed on both original speech and the same speech with added white noise. For each case, constant, linear and quadratic time-varying frequency models were used. We also recorded computation time to compare computational efficiency.

Typical results for clean data (no additive noise) are shown in figure 4 and figure 5. In figure 4, the upper panel depicts comparison of pitch estimation results and hand marked pitches; the lower panel depicts the HNR and the subsequent result for U/V classification. The U/V discrimination threshold was set at 3 dB by averaging results of other utterances. In figure 5, we can clearly see that the HNR performs very well as an indication of U/V classification.

The comparisons for different noise levels and frequency models are shown in table 2. We used a previously-defined relative accuracy [8] to qualify the performance of the algorithm, which is defined as

$$\text{Relative Accuracy} = \left[1 - \frac{1}{N}\sum_{k=1}^{N}\frac{|f(k) - f_{ELSH}(k)|}{f(k)}\right] \times 100\% , \qquad (16)$$

where $f(k)$ and $f_{\text{ELSH}}(k)$ are the hand marked pitch and estimated pitch from ELSH model at k^{th} pitch period, respectively. The accuracy values show that the ELSH approach performs extremely well in low SNR situations. Going from clean data to noise level of -10dB, there is only about a 1.3 percent drop in overall accuracy. It should also be noted that best results were obtained by using different window size for different noise levels.

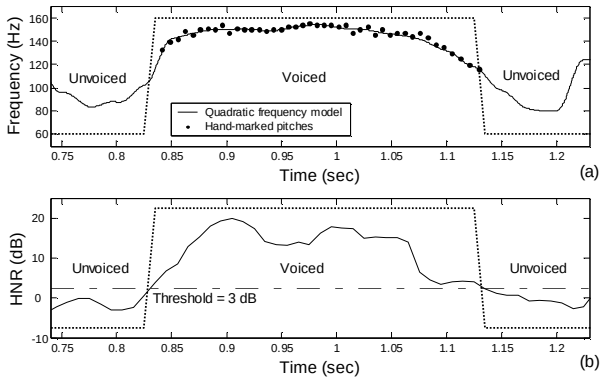

Figure 4. Results of pitch estimation (no additive noise). (a) shows results of linear frequency model and hand-marked pitches. (b) shows HNR (solid line) and decisions of U/V classification (dotted line, high level – voiced; low level – unvoiced) . The threshold is 3 dB.

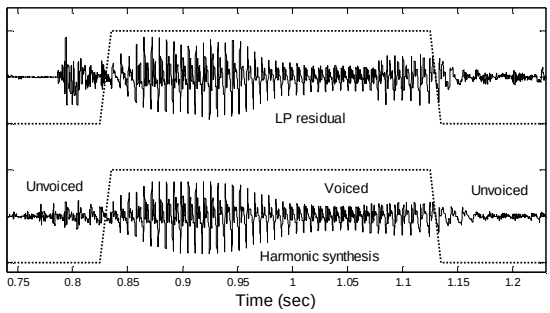

Figure 5. LP residual and Harmonic Synthesis (no additive noise). The upper trace is residual signal and lower trace is harmonic synthesis. The dotted line separates estimated voiced and unvoiced regions.

Frequency Model	Constant	Linear	Quadratic
No additive noise 50 ms data window	98.9%	99.1%	99.2%
White noise (SNR= -5 dB) 50 ms data window	98.4%	98.6%	98. 7%
White noise (SNR= -10 dB) 100 ms data window	98.0%	97.8%	97.9%
White noise (SNR= -15 dB) 100 ms data window	92.7%	92.6%	92.6%
Approx. Computation Time (sec)	110	340	830

Table 2. Relative accuracy for different noise levels and frequency models. The algorithm is implemented in MATLAB R13 running on AMD Athlon 1.5GHz system with Windows 2000. The computation time is for 5 second speech sampled at 16 KHz.

There are differences between frequency models. At low noise level (SNR≥ -5dB), a higher order model generates better results, but also takes more computation time; at high noise level (SNR≤ -10dB), a higher order model tends to fit noise and thus generates worse results. In practice, a proper order should be chosen by balancing between accuracy and computation efficiency. Order is also affected by data quality and window size. In general, a short window potentially has less frequency variation and a low order model may be sufficient; a long window potentially has more frequency variation and thus may need a high order model, yet offers higher tolerance to noise. For the speech tests, noise levels, and window sizes presented in this paper, since the advantages of the quadratic model were slight, we preferred the linear frequency model at low noise level.

5. CONCLUSION

In this paper, we presented an approach for harmonic analysis, which is an extension to LSH. It has been demonstrated that this extended LSH approach not only is exceptionally robust and accurate at low SNR, but also is capable of capturing rapid frequency change. Two applications of this approach were shown in the paper. The application to acoustic heartbeat detection, where the quadratic time-varying model was used, has shown success on difficult data. The application to pitch estimation has potential for high resolution pitch estimation. The disadvantage of this method is that the computation complexity is high. It is thus not currently efficient to use this approach for U/V classification only. Future work on a large speech data base is needed to confirm the conclusion of sufficiency of the linear time-varying model.

We acknowledge help from technical discussions with Prof. Mari Ostendorf. This work was funded by the Army Research Lab.

6. REFERENCES

[1] N. Abu-Shikhah and M. Deriche, "A robust technique for harmonic analysis of speech," in *Proc. IEEE ICASSP'01*, vol. 2, pp. 877-80, Piscataway, NJ, 2001.

[2] Ahn R, Holmes Wh, Deriche M, Moody M, and Bennamoun M, "Harmonic-plus-noise decomposition and its application in voiced/unvoiced classification," in *Proc. IEEE TENCON '97*, vol. 2, pp. 587-90, Brisbane, Australia, 1997.

[3] D. W. Griffith and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1223-1235, 1988.

[4] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, pp. 744-754, 1986.

[5] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. IEEE ICASSP'90*, vol. 1, pp. 249-252, 1990.

[6] R. J. McAulay and T. F. Quatieri, "The sinusoidal transform coder at 2400 b/s," in *IEEE MILCOM '92*, vol. 1, pp. 378-380, 1992.

[7] M. Scanlon, "Acoustic monitoring of first responder's physiology for health and performance surveillance," in *SPIE 16th Annual International Symposium on Aerospace/Defense Sensing, Simulation, and Controls*, Orlando, Florida, USA, 2002.

[8] A. Shah, R. P. Ramachandran, and M. A. Lewis, "Robust pitch estimation using an event based adaptive gaussian derivative filter," in *IEEE ISCAS'02*, vol. 2, pp. 843-846, 2002.