

# A NEW APPROACH TO HARMONIC ANALYSIS OF SPEECH

*Nazih Abu-Shikhah*  
n.abushikhah@qut.edu.au

*Mohamed Deriche*  
m.deriche@qut.edu.au

Signal Processing Research Centre  
School of Electrical & Electronic Systems Engineering  
Queensland University of Technology  
GPO Box 2434, Brisbane, Q4001, Australia

## ABSTRACT

In this paper we present a new approach, called Linear Harmonic Analysis (LHA), to compute the harmonic component of a given speech signal. The technique is based on the harmonic model of speech which is derived for a given range of fundamental frequencies using a set of linear equations. The fundamental frequency resulting in the minimum Mean Square Error (MSE) is selected as the target harmonic frequency, and hence the harmonic component of the signal is estimated. The method was applied to different analytical and real speech signals at different levels of Signal to Noise Ratio (SNR). Results showed that LHA method outperforms the Total Least Square Prony method (TLSP), and produces reliable estimates at low SNRs.

Keywords: *Harmonic, Fundamental frequency, Total Least Squares Prony*

## 1. INTRODUCTION

A speech signal can be modelled using two components: a quasi-periodic part ( Harmonic component) and a non-periodic part (Noise like component). This is referred to as Harmonic + Noise (H+N) model, and is widely used in speech enhancement, recognition, synthesis, and coding. In speech coding, implementation of this model is found in Multiband Excitation Coding (MBE) [1], sinusoidal coding [2], and harmonic coding [3]. The optimal estimate of the harmonic component (and hence noise component) is the most crucial

step for such coders. Methods based on Discrete Fourier Transform are widely implemented, but these may result in an inaccurate estimate of the fundamental frequency. This is highly pronounced when the signal is corrupted with noise. Another widely implemented method is the Total Least Squares Prony (TLSP)[4]. This technique produces good results for clean signals, however at low SNR, it suffers from the same problems as the DFT.

Our aim here is to develop a more robust harmonic analysis that results in an optimum estimate for the fundamental frequency, and hence an optimal harmonic component of the input signal. This is achieved, by modelling input speech as a sum of harmonics, and estimating the periodic component using a set of linear equations model. This set of linear equations is derived from minimising mean square of error (MSE) between the input and estimated signal. The process needs to be repeated for all possible values (extensive search) of the fundamental frequency of speech ranging between 50 and 400 Hz. Once the MSE for different frequencies are computed, the harmonic component is selected as the one obtained from the minimum MSE. The model proved to give good results that outperform both DFT and TLSP methods, particularly when corrupted signals are processed. The major disadvantage of the method is the large computation load required to perform the extensive search over the range of potential fundamental frequencies. Reduction of search time can be carried out by computing a rough estimate of the fundamental frequency fol-

lowed by an extensive search within the vicinity of this estimate. The rough estimate of the fundamental frequency can be obtained using the signal spectrum, or by using one of the pitch period estimation techniques[5], e.g. Autocorrelation based, or Cepstral based methods. This results in a considerable savings in computational time.

This paper is organised as follows: section 1 describes the H+N model, section 2 presents the derivation for the proposed algorithm which we refer to as LHA (Linear Harmonic Analysis). In section 3 experimental results of the LHA algorithm for both clean and noisy signals is compared to that of DFT and TLSP, then a conclusion is presented.

## 2. HARMONIC+NOISE DECOMPOSITION

The Harmonic plus Noise (H+N) model assumes that each frame of speech is composed of sum of sinusoids, located at multiples of the fundamental frequency, in addition to a non-harmonic signal considered to be a noise like component of speech. Due to the quasi-stationary nature of speech, the model parameters are assumed to be time invariant, and thus the harmonic is given by:

$$\begin{aligned} \mathbf{s}(k) &= \sum_{i=1}^P A_i \cos(2\pi i \frac{f_0}{f_s} k) + \mathbf{n}(k) \\ &= \mathbf{h}(k) + \mathbf{n}(k) \end{aligned} \quad (1)$$

where,  $\mathbf{s}(k)$  is the speech signal of length  $L$ , with its two components:  $\mathbf{h}(k)$ , and  $\mathbf{n}(k)$  being the harmonic, and noise components, respectively.  $P$  is the total number of harmonics,  $A_i$  is the amplitude for harmonic number  $i = 1, 2, \dots, P$ ,  $f_0$  is the fundamental frequency,  $f_s$  is the sampling frequency, and  $k = 0, 1, \dots, L - 1$  is the time index.

## 3. PROPOSED LINEAR HARMONIC ANALYSIS (LHA) ALGORITHM

To derive our optimisation algorithm, we start by writing the expansion for the mean of squares of errors (MSE) between  $\mathbf{s}(k)$  and  $\mathbf{h}(k)$  given as:

$$\begin{aligned} E &= \sum_{k=0}^{L-1} [\mathbf{s}(k) - \mathbf{h}(k)]^2 \\ &= \sum_{k=0}^{L-1} \left[ \mathbf{s}(k) - \sum_{i=1}^P A_i \cos(2\pi i \frac{f_0}{f_s} k) \right]^2, \quad (2) \\ & \quad i = 1, 2, \dots, P \end{aligned}$$

Minimising MSE involves solving the following equations:

$$\begin{aligned} F_i &= \frac{\partial E}{\partial A_i} \\ &= \sum_{k=0}^{L-1} \left\{ \left[ \mathbf{s}(k) - \sum_{i=1}^P A_i \cos(2\pi i \frac{f_0}{f_s} k) \right] \cos(2\pi i \frac{f_0}{f_s} k) \right\} \\ &= 0, \quad i = 1, 2, \dots, P \end{aligned} \quad (3)$$

Hence, we have  $P$  functions expressed in terms of our model parameters. rearranging (3) results in the following set of equations expressed in a matrix form:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \cdots & \vdots \\ x_{P1} & x_{P2} & \cdots & x_{PP} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_P \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_P \end{bmatrix}. \quad (4)$$

or in a compact form as:

$$\mathbf{XA} = \mathbf{Y} \quad (5)$$

where,

$$x_{ij} = \sum_{k=0}^{L-1} \cos(2\pi i \frac{f_0}{f_s} k) \cos(2\pi j \frac{f_0}{f_s} k),$$

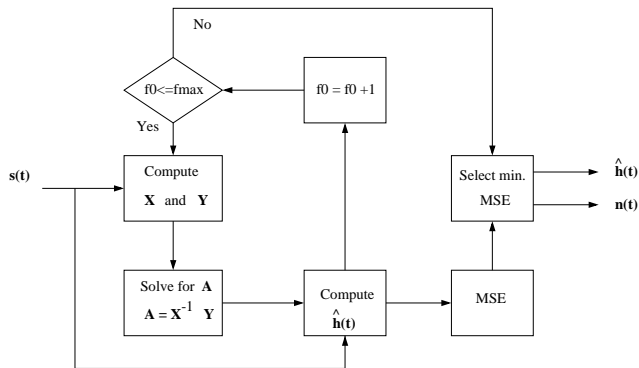


Figure 1: LHA Algorithm

$$y_i = \sum_{k=0}^{L-1} \mathbf{s}(k) \cos(2\pi i \frac{f_0}{f_s} k), \text{ and}$$

$$i, j = 1, 2, \dots, P.$$

The harmonic amplitudes  $\mathbf{A}$  are then expressed as:

$$\mathbf{A} = \mathbf{X}^{-1} \mathbf{Y} \quad (6)$$

The procedure for finding the optimum set of  $\mathbf{A}$ 's is shown in figure(1), and is implemented as follows:

1. Given the range of possible fundamental frequencies, which is between 50-400 Hz for speech signals, start with the minimum frequency in that range, and form equation(4).
2. For the selected frequency, compute harmonic amplitudes  $\mathbf{A}$  using equation(6).
3. Find the estimated harmonic component of the speech signal using (1).
4. Compute the MSE for that estimate, using(1).
5. Increment the frequency by 1, and repeat steps 1-4 for all frequencies in the range.
6. Select the frequency and harmonic amplitudes that result in the minimum MSE.

It should be noted that the method involves an extensive search through all possible fundamental frequency values. The search time can be considerably reduced by first taking the Discrete Cosine

Transform (or alternatively the Discrete Fourier Transform) of  $\mathbf{s}(k)$ , the range of values can be reduced to only selecting frequencies in the vicinity of a selected number of peaks in the spectrum. The LHA can hence be applied with much less computations. Alternatively, a rough estimate of the fundamental frequency can be obtained using one of the pitch period estimation techniques, e.g. Autocorrelation based, or Cepstral based methods. This is then followed by an extensive search within the vicinity of this estimate.

#### 4. PERFORMANCE

The performance of our LHA algorithm was tested using two analytical signals. Its performance was also compared to the that of the TLSP method as described below.

A periodic signal  $\mathbf{s}$  of length  $L=200$  samples, figure(2), was generated using equation(2) with:

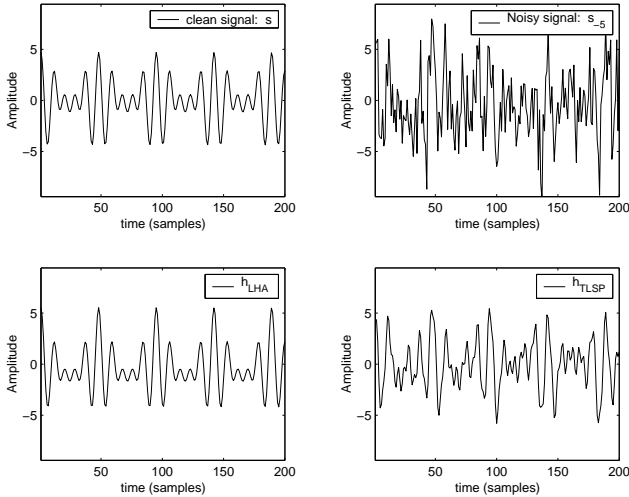
$$P=5, f_0=170 \text{ Hz}, F_s= 8000 \text{ Hz}, \text{ and} \\ \mathbf{A}_{1,2,\dots,5}=[-0.0158,-0.2827,0.9481,2.0677,1.9873]$$

Corrupted versions of  $\mathbf{s}$ , referred to as  $\mathbf{s}_{-5}$ ,  $\mathbf{s}_0$ , and  $\mathbf{s}_{10}$  were generated by adding random noise with signal to noise ratios of -5 dB, 0 dB, and 10 dB, respectively. All three signals were processed using LAH and TLSP methods, and each method results in a harmonic component ( $\mathbf{h}_{LHA}, \mathbf{h}_{TLSP}$ ), and noise components ( $\mathbf{n}_{LHA}, \mathbf{n}_{TLSP}$ ). The MSEs between the original signals and estimated harmonic components were computed. 100 simulations were performed for each of the SNR levels, using different randomly generated noise signals in each simulation. The average MSE at each SNR was computed, and the results are summarised in table(1).

Table(1), shows that for high SNR both LHA and TLSP leads to approximately the same MSE, however, for lower SNR, LHA method results in a far better estimate of the harmonic component than the TLSP method. The comparison was implemented for other SNR levels, and the results obtained showed that LHA method, outperforms

SNR dB	Average MSE for $\mathbf{h}_{LHA}$	Average MSE for $\mathbf{h}_{TLSP}$
10	0.006	0.051
0	0.07	0.43
-5	0.22	1.535

Table 1: MSE for different SNR levels

Figure 2: Comparison of  $\mathbf{s}$ , noisy  $\mathbf{s}$  and estimated periodic component at SNR=-5 dB

TLSP, particularly at lower SNR values, which is clear from the above table.

Figure(2) illustrates a comparison of  $\mathbf{s}$  and the periodic components obtained using both methods for one simulation at SNR=-5 dB, the estimated parameters compared to the original values are also shown in table(2). It is obvious that the estimated parameters using LHA are very close to original values, while the ones estimated using TLSP are far from the original values.

The above procedure was implemented on a real speech segment of length 240 samples. The sampling frequency is 8 kHz, and the segment is voiced with pitch period = 32 samples (i.e. 250 Hz) computed using autocorrelation function method [5]. This implies that the total number of harmonics for this signal is  $\lfloor \frac{4000}{250} \rfloor = 16$ , where,  $\lfloor \cdot \rfloor$  is the lower nearest integer. The signal was processed using LHA and TLSP and the results are shown in figure(3). Bearing in mind that we are mod-

Parameters	Original	LHA	TLSP
A1	-0.0158	-0.2750	1.2630
A2	-0.2827	-0.0056	3.8777
A3	0.9481	1.1479	0.0234
A4	2.0677	1.9562	0.0005
A5	1.9873	2.2301	0.0008
f1 (Hz)	170	170	512
f2 (Hz)	340	340	675
f3 (Hz)	510	510	680
f4 (Hz)	680	680	845
f5 (Hz)	850	850	2383

Table 2: Estimated Parameters

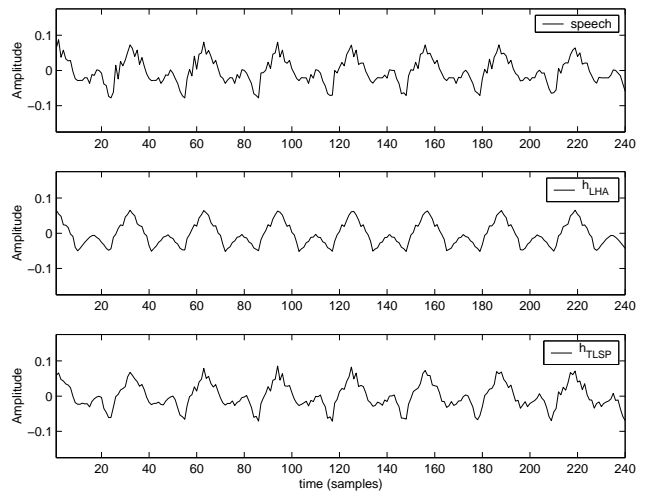


Figure 3: Comparison of speech segment and estimated periodic components

elling speech as sum of harmonics, we would expect that TLSP will try to find the best fit to speech, with the best frequency pattern that is not necessarily a harmonic model. This fact can be seen in the above figure, where  $\mathbf{h}_{TLSP}$  fits  $\mathbf{s}$  better than  $\mathbf{h}_{LHA}$ . However, the estimated fundamental frequency using LHA was 257 Hz (very close to 250 Hz), while the minimum frequency obtained from TLPS was 113 Hz (about half 250 Hz). In summary, it is evident that LHA is capable of estimating the actual fundamental frequency and corresponding harmonic component of speech, while TLSP finds the best fit to the given signal. Figure(4) shows a spectrum comparison of the above signals.

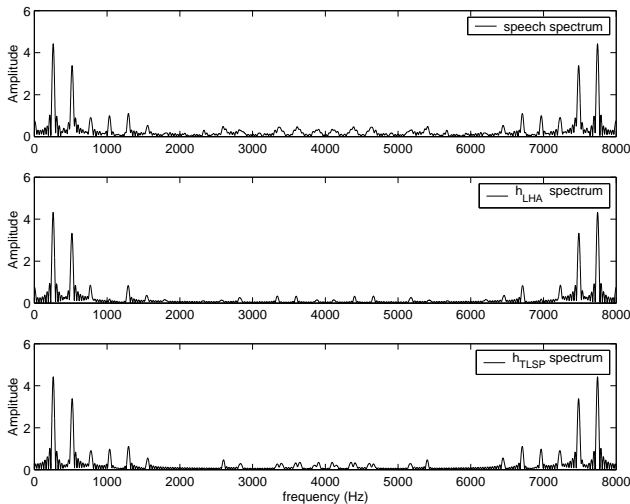


Figure 4: Comparison of spectrums of speech segment and estimated periodic components

## 5. CONCLUSION

We presented a novel technique for decomposing speech signal into harmonic and noise like components which can be used in different fields of speech applications. The method implemented is based on solving a set of linear system of equations for the expected fundamental frequency range of speech (50-400 Hz). Experiments on analytical, and speech signals, showed the superiority of the method over the widely implemented TLPS method. Tests for different levels of SNR, proved that LHA outperforms TLSP. At low SNR levels, whilst TLSP was unable to result in good estimates signal fundamental frequency and corresponding harmonic amplitudes, the LHA showed a much better performance. The method major disadvantage is its high computation load due extensive search. This can be improved significantly by using an initial estimate of the fundamental frequency using the spectrum of the input signal, or using pitch period estimation methods such as the autocorrelation based technique.

## 6. REFERENCES

- [1] D. Griffin and J.S. Lim. Multiband excitation vocoder. *IEEE Transaction on Acoustics, Speech, and Signal Processing*, 36(8):1223–1235, August 1988.
- [2] R. McAulay and T.F. Quatieri. Speech analysis/synthesis based on sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, August 1986.
- [3] G. Yang and H. Leich. High quality harmonic coding at very low bit rates. *ICASSP'94*, 1:181–184, 1994.
- [4] M. Rahman and K. Yu. Total least squares approach for frequency estimation using linear prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10):1440–1454, October 1987.
- [5] W.B. Kleijn and K.K. Paliwal. *Speech Coding and Synthesis*. Elsevier Science B.V., 1995.