A ROBUST TECHNIQUE FOR 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

A technique named Least Squares Harmonic (LSH) is proposed for speech decomposition. The problem of harmonic estimation for speech is formulated as a solution to two sets of linear equations derived from minimising the Mean Squared Error between original and estimated signals. The algorithm assumes that a good initial estimate of the pitch period is available. The performance of the algorithm is comparable to that of the Total Least Square Prony method (TLSP) at high Signal-to-Noise ratios, however, at very low SNR, the proposed algorithm leads to a much more accurate harmonic representation. The approach used here has a great potential in coding and recognition applications.

Keywords: Least Squares, Harmonic

1. INTRODUCTION

In the Harmonic plus Noise (H+N) model, speech is decomposed into two components: a quasi-periodic part (Harmonic component) and a non-periodic part (Noise-like component). This model is widely used in speech enhancement, recognition, synthesis, and coding. For instance, Multiband Excitation (MBE) [1], sinusoidal [2], and harmonic coding [3] are all based on the above model. The crucial step in such applications is to correctly identify an optimal harmonic component (and hence noise component). A widely used method for harmonic estimation is the Total Least Squares Prony (TLSP)[4]. This technique produces good results for clean signals, however its performance deteriorates rapidly at low SNR.

Our objective here is to develop a robust harmonic analysis algorithm with an accurate estimate of the fundamental frequency. We extend our previous work[5] by including the phase information in the model. For a given fundamental frequency, the algorithm leads to solving two sets of linear equations derived from minimising the mean square error (MSE) between input and estimated signals. For each fundamental frequency *candidate* in the range 50 and 400 Hz, the process is repeated (exhaustive search) and a set of MSEs is computed. The harmonic component is selected as the one resulting in the minimum value of the MSE. The technique requires large computational load due to the extensive search over the range of potential fundamental frequencies. Computation time, however, is reduced considerably by computing a rough estimate of the fundamental frequency followed by the above extensive search in the vicinity of this estimate. The rough estimate of the fundamental frequency can be obtained using signal spectrum, or by using one of the pitch period estimation techniques [6], e.g. Autocorrelation, or Cepstral methods.

This paper is organised as follows: section 2 describes the H+N model, section 3 presents the derivation for the proposed algorithm. In section 4, experimental results of the algorithm for both analytical and real speech signals are compared to those of TLSP, then a conclusion is presented in section 5.

2. THE SINUSOIDAL AND HARMONIC+NOISE MODELS

The sinusoidal model developed in [2] is based on passing an excitation signal for each frame through a linear time-varying filter. For any particular speech segment (frame), the excitation signal is assumed to consist of a set of sinusoids and the resulting speech signal is hence

given by eqn(1):

$$\begin{aligned} s(k) &= \sum_{i=1}^{M} C_i cos(2\pi i \frac{f_i}{f_s} k + \phi_i) \\ &= \sum_{i=1}^{M} C_i cos(i\omega_i k + \phi_i) \end{aligned} \quad (1)$$

where $s(k)$ is the speech segment of length N, M is the total number of sinusoids in $s(k$, C_i, ,f_i, and ϕ_i are the amplitude, frequency and phase angle of each sinusoid $i = 1, 2, ...M$, f_s is the sampling frequency, $\omega_i = 2\pi \frac{f_i}{f_s}$ is the normalised frequency, and $k = 0, 1, ..., N-1$ is the time index. The Harmonic plus Noise (H+N) model, on the other hand, is a special case of the sinusoidal model where a frame of speech is assumed to be composed of a sum of sinusoids, located at multiples of the fundamental frequency, in addition to a non-harmonic signal considered to be noise-like. For a short speech signal, the model parameters are assumed to be time invariant, due to the quasi-stationary nature of speech. The (H+N) model is given by:

$$\begin{aligned} s(k) &= \sum_{i=1}^{P} C_i cos(2\pi i \frac{f_0}{f_s} k + \phi_i) + n(k) \\ &= \sum_{i=1}^{P} C_i cos(i\omega_0 k + \phi_i) + n(k) \\ &= h(k) + n(k) \end{aligned} \quad (2)$$

where $h(k)$,and $n(k)$ are the speech harmonic, and noise components, respectively. P is the total number of harmonics, C_i, and ϕ_i are the amplitude and phase angle corresponding to harmonic $i = 1, 2, ...P$, f_0 is the fundamental frequency,f_s is the sampling frequency, $\omega_0 = 2\pi \frac{f_0}{f_s}$ is the normalised fundamental frequency, and $k = 0, 1, ..., N-1$ is the time index.

It is evident, from the above, that for the sinusoidal model, the larger the number of sinusoids M is, the closer the reconstructed signal is to the original signal. In contrast, the H+N model, uses only P sinusoids, where P is less than M, to represent the harmonic part and the remaining $M - P$ sinusoids are summed together to represent the noise part of the signal. This makes the H+N model model attractive in a number of speech applications.

3. PROPOSED TOTAL LEAST SQUARES HARMONIC(LSH) ALGORITHM

The proposed algorithm is based on the H+N model. It starts by rewriting the harmonic component given in equation(2) as:

$$\begin{aligned} h(k) &= \sum_{i=1}^{P} C_i [cos(i\omega_0 k) cos(\phi_i(k)) \\ &\quad - sin(i\omega_0 k) sin(\phi_i(k))] \\ &= \sum_{i=1}^{P} A_i cos(i\omega_0 k) - B_i sin(i\omega_0 k) \end{aligned} \quad (3)$$

where, $A_i = C_i cos(\phi_i)$, $B_i = C_i sin(\phi_i)$. It should be apparent that $\phi_i = tan^{-1}(\frac{B_i}{A_i})$.

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

$$\begin{aligned} E &= \frac{1}{N} \sum_{k=0}^{N-1} [s(k) - h(k)]^2 \\ &= \frac{1}{N} \sum_{k=0}^{N-1} [s(k) - \sum_{i=1}^{P} A_i cos(i\omega_0 k) + B_i sin(i\omega_0 k)]^2, \end{aligned} \quad (4)$$

For a given fundamental frequency, f_0, the minimum MSE is found by setting:

$$\frac{\partial E}{\partial A_j} = 0, \quad and \quad \frac{\partial E}{\partial B_j} = 0, \quad for \quad j = 1, 2, \ldots, P \quad (5)$$

After algebraic manipulations, the above equations can be expressed in matrix form:

$$\mathbf{Y_1} = \mathbf{QA} + \mathbf{RB} \quad (6)$$

$$\mathbf{Y_2} = \mathbf{SA} + \mathbf{TB} \quad (7)$$

where $\mathbf{A}$ and $\mathbf{B}$ are Px1 unknown vectors to be estimated, and the remaining matrices are defined as:

$\mathbf{Q}(i,j) = \sum_k cos(i\omega_0 k) cos(j\omega_0 k)$,
$\mathbf{R}(i,j) = -\sum_k sin(i\omega_0 k) cos(j\omega_0 k)$,
$\mathbf{S}(i,j) = \sum_k cos(i\omega_0 k) sin(j\omega_0 k)$,
$\mathbf{T}(i,j) = -\sum_k sin(i\omega_0 k) sin(j\omega_0 k)$,
$\mathbf{Y_1}(j) = \sum_k s(k) cos(j\omega_0 k)$,
$\mathbf{Y_2}(j) = \sum_k s(k) sin(j\omega_0 k)$
where $i, j = 1, 2, ..., P$, and $k = 0, 1, ..., N-1$.

Solving equations 5 and 6 results in:

$$\mathbf{A} = (\mathbf{S} - \mathbf{TR^{-1}Q})^{-1}(\mathbf{Y_2} - \mathbf{TR^{-1}Y_1}) \quad (8)$$

$$\mathbf{B} = \mathbf{R^{-1}}(\mathbf{Y_1} - \mathbf{QA}) \quad (9)$$

The procedure for finding $\mathbf{A}$, and $\mathbf{B}$ is shown in figure(1):

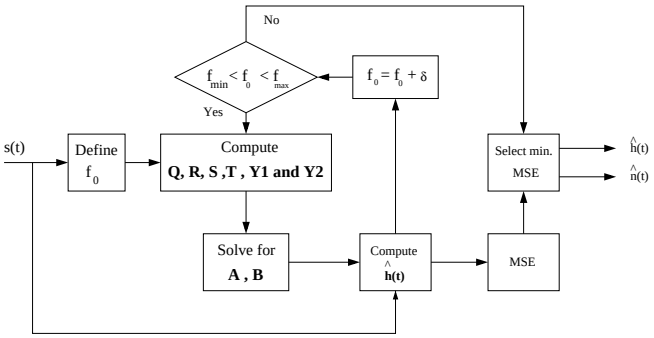

Figure 1: LSH Algorithm

1. Start with an estimate of the fundamental frequency f_0, and define a frequency range around it.
2. Solve for harmonic amplitude vectors **A** , and **B** using equ.(8,9).
3. Find the estimated harmonic component of the speech signal using equ.(3).
4. Compute the MSE for that estimate, using equ.(4).
5. Increment the frequency by δ, and repeat steps 1–4 for all frequencies in the range of interest.
6. Select the frequency and harmonic vectors that result in the minimum MSE.

4. PERFORMANCE ANALYSIS

The performance of the proposed LSH algorithm was compared to that of the TLSP using an analytically generated signal, and a real speech signal.

4.1. Analytical signal

We generated a periodic signal $\mathbf{s_{or}}$ of length N=200 using equ.(3) with parameters: P=10, f_0=203 Hz, f_s= 8000 Hz. The first few parameters are given in table (2). Noisy observations of $\mathbf{s_{or}}$, at different SNRs were then obtained. All signals were processed using the LSH and the TLSP methods, and MSE over 100 realisations was computed at each SNR level. The results are summarised in table(1).

The results clearly show that the LSH method, outperforms the TLSP for all SNR levels. The LSH continues to perform extremely well even a very low SNR. Table(2) gives the first five (out of ten) values for the original and estimated amplitudes and phases, based on the two methods at SNR = 0 dB. It is obvious that the estimated parameters using LSH are very close to the true values, in comparison to the TLSP. Numerous other experiments have been carried, all showing the that the proposed technique is more robust in the presence of noise compared to the TLSP.

SNR dB	**Average MSE for $\mathbf{h}_{LSH}$**	**Average MSE for $\mathbf{h}_{TLSP}$**
10	0.028	0.071
0	0.24	0.73
-5	0.81	2.56
-10	3.1	7.98
-15	12.5	24.4

Table 1: MSE for different SNR levels

A_{or}	A_{LSH}	A_{TLSP}	ϕ_{or}	ϕ_{LSH}	ϕ_{TLSP}
1.40	1.38	1.38	1.36	1.28	0.72
0.70	-0.66	0.58	-3.50	-0.59	1.25
-0.39	-0.37	0.51	-0.05	-0.40	1.91
-0.49	-0.47	0.51	0.32	0.51	3.04
-1.69	1.71	1.67	-2.30	0.78	0.93

Table 2: First 5 true and estimated parameters at SNR=0 dB

4.2. Real speech signal

The proposed algorithm was then used on a real speech segment of length 256 samples. The sampling frequency was 8 kHz, and the segment is voiced with pitch period = 31 samples (i.e. 258 Hz) computed using the autocorrelation method [6]. This implies that for a fundamental frequency around 258 Hz, the number of harmonics is expected to be between 15 and 17. The signal was processed using the LSH and the TLSP and the comparison between a segment of the original and estimated signals illustrated in figure(2). The "fine-tuned" fundamental frequency estimated using the LSH was found to be 263 Hz, while the estimated frequency using the TLSP was found to be 265 Hz. Bearing in mind that the TLSP assumes that the given speech data consists of a harmonic model only, we would expect it to give a better fit to the harmonic structure in the data. However, we found that the TLSP missed completely several harmonics (see table (3)). The table shows that, in the TLSP, harmonics (4,5,6,7,and 15) are completely missed. A comparison of the spectra from original and estimated signals is shown in figure(3).

LSH	$TLSP$	LSH	$TLSP$
1	1	10	10.04
2	1.99	11	10.94
3	3.01	-	11.14
4	-	12	12.12
5	-	-	12.8-
6	-	13	13.05
7	-	-	13.29
8	7.91	14	14.1
9	9.07	15	-
-	9.32		

Table 3: Harmonics present in the signal

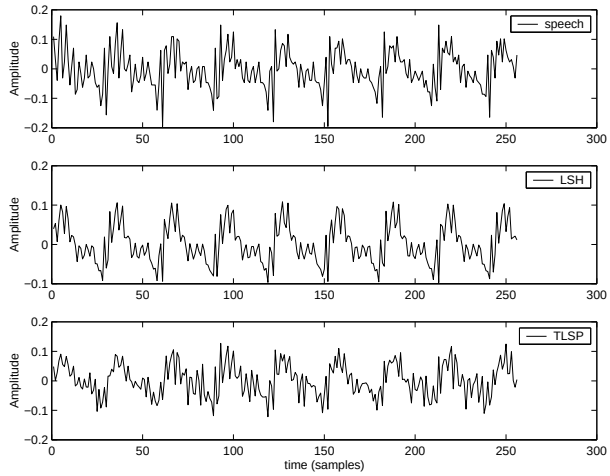

Figure 2: Comparison of speech segment and estimated periodic components

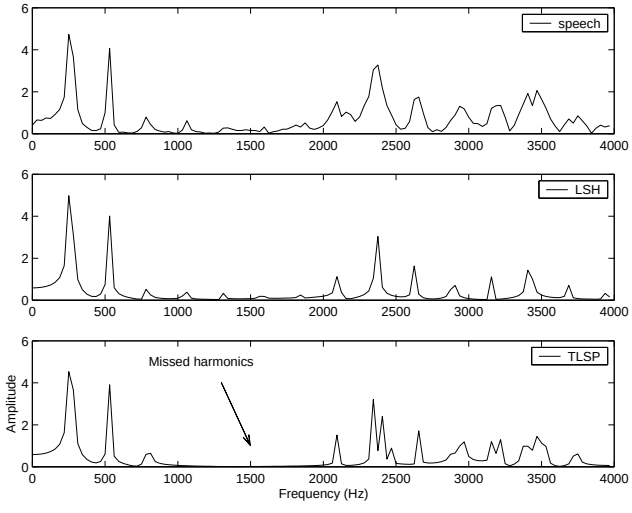

Figure 3: Comparison of spectra of speech segment and estimated periodic components

5. CONCLUSION

We presented here a novel technique for decomposing speech signal into harmonic and noise-like components. The method implemented is based on solving two sets 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. The extensive tests carried for different levels of SNR, proved that LSH outperforms TLSP. At low SNR levels, whilst TLSP was unable to result in a good estimate for the signal fundamental frequency and corresponding harmonic amplitudes, the LSH displayed a much better performance . Whilst the method requires high computation load due extensive search, its performance has been significantly improved using an initial estimate of the fundamental frequency obtained from a pitch period estimation method such as the autocorrelation, or the cepstrum based technique. We intend to implement this method in a low bit rate harmonic coder operating at around 2.4 kb/s and aim at achieving high quality in synthesised speech.

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] N. Abu-Shikhah and M. Deriche. A new approah to harmonic analysis of speech. *Proceedings of DSP 2000*, October 2000, Texas, USA.

[6] W.B. Kleijn and K.K. Paliwal. *Speech Coding and Synthesis*. Elsevier Science B.V., 1995.