

# SINGLE COMPLEX SINUSOID AND ARHE MODEL BASED PITCH EXTRACTORS

*Ilija Zeljkovic and Yannis Stylianou*

AT&T Labs-Research, Shannon Laboratories, 180 Park Ave, Florham Park, NJ 07932-0971  
{ilija, yannis} @research.att.com

## ABSTRACT

In this paper we propose two techniques for the estimation of the fundamental frequency of speech signals. The first technique is based on the Autoregressive Harmonic Excitation (ARHE) speech model. ARHE model consists of an autoregressive process driven simultaneously by white noise and a periodic excitation. The second technique is based on the estimation of a complex sinusoid in white Gaussian noise. It uses the Hilbert transform of the speech signal and the derivative of its phase function over the time. The derivative of the phase information is seen as a simple model of a moving average process driven by noise. The fundamental frequency is obtained by the minimum variance estimator of the model. The proposed methods have comparable performance to previous reported pitch detectors while they maintain their performance under noisy conditions.

## 1. INTRODUCTION

The estimation of the fundamental frequency has engaged a lot of effort in the speech analysis community. Its reliable estimation is of high importance and has many applications to speech coding, speech synthesis and speech recognition. Many algorithms for the estimation of fundamental frequency have been proposed in the literature. In [1], a wide collection of time and frequency domain techniques are presented. In [2] the fundamental frequency is estimated by seeking the minimum of the mean-squared error (MSE) between a sinusoidal representation of the original speech waveform and its harmonic representation. In [3] the fundamental frequency is determined from the spacing between peaks in a selected portion of the spectrum. In [4] the fundamental frequency is estimated by flattening the spectrum of the signal by a bank of bandpass lifters and extracting the pitch frequency from autocorrelation functions calculated at the output of the lifters. In [5] a normalized cross-correlation criterion is used along with a strategy of dynamic programming and post filtering in order to avoid doubling and halving of the pitch frequency. Finally, in [6] another cross-correlation criterion was used based on the work initially proposed in [7].

Although of the above list of the proposed techniques, accurate and robust, under noise conditions, pitch detection remains an open problem. In this paper we propose two new algorithms for the estimation of the fundamental

frequency of speech signals. The first pitch estimator is based on the recently proposed Autoregressive Harmonic Excitation (ARHE) Speech Model[8] and the second on the estimation of a complex sinusoid in white Gaussian noise. The ARHE model consists of an autoregressive (AR) process driven simultaneously by white noise and a periodic excitation that is modeled by its Fourier expansion into fundamental and higher harmonics. The AR parameters, harmonic amplitudes, and unknown excitation noise intensity are estimated by means of Total Least Squares (TLS) and Singular Value Decomposition (SVD). The TLS technique employed is based on rank reduction of the combined speech and the excitation signal data matrix by setting the smallest singular value of the matrix to zero. The fundamental frequency,  $\omega_0$ , of the excitation function is varied in a prescribed range and the smallest singular value of the corresponding data matrix is computed. The frequency that produces the minimum of all of the smallest singular values is chosen as a fundamental frequency.

The second technique is based on the estimation of the frequency of a single complex sinusoid in white Gaussian noise. The method proposed here uses the Hilbert transform of the speech signal and the derivative of the phase function over the time. It can be shown that the Maximum Likelihood Estimator (MLE) of the fundamental frequency is given by a weighted sum of the differenced phase data (see Section 3). There is an advantage of using differenced phase data since it avoids phase unwrapping. Additionally, using the Hilbert transform of the speech signal, the optimum MLE may be written as a weighted sum of a correlation of the signal. This leads to a very fast algorithm for the estimation of the fundamental frequency avoiding any FFT implementation while it is much simpler and more robust for moderate SNRs than the estimators based on the periodogram.

The above two algorithms are compared against two previously reported pitch detectors[6],[5]. The performance of the proposed algorithms is evaluated on clean speech as well as on low Signal to Noise Ratios (SNR's) using a weighted Gross Pitch Error (GPE) criterion [9]. Results show that the proposed methods provides similar performance to the reference pitch detectors ([6],[5]) and their performance is maintained at low SNR. The estimated pitch values from these four pitch detectors were also used in modeling the speech signal as a sum of harmonics[10] in voiced only areas. The voiced/unvoiced decisions are made by the algorithm proposed in[5]. In all cases, high

quality of reconstructed speech was obtained.

The paper is organized as follows. Section 2 presents the ARHE model and its application to the estimation of fundamental frequency, This is followed in Section 3 by the description of the method based on the differenced phase data of the Hilbert transform of the speech signal. Section 4 presents results of applying the proposed methods for the estimation of fundamental frequency in a large speech database in order to support our conclusions.

## 2. THE ARHE MODEL

The ARHE model assumes that the vocal tract is excited by both the glottal excitation, which is periodic, and white-noise-like excitation. In the ARHE model the glottal excitation is approximated by a Fourier series expansion with fundamental frequency  $\omega_0$ , as shown in Fig. 1. Then, the difference equation that governs speech produc-

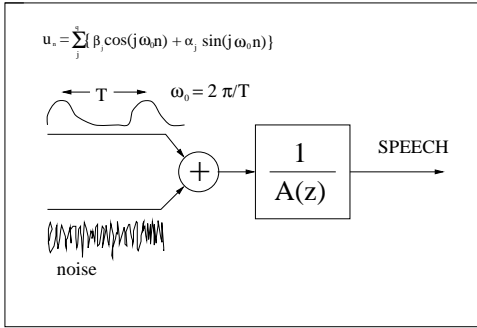


Figure 1: ARHE Speech Production Model

tion process becomes:

$$y_n = - \sum_{k=1}^p a_k y_{n-k} - \sum_{j=0}^q (\beta_j \cos j\omega_0 n + \alpha_j \sin j\omega_0 n) + e_n \quad (1)$$

where  $0 \leq n < N$ . The coefficients  $a_k$  represent Linear Prediction coefficients and  $\beta_j$  and  $\alpha_j$  define unknown amplitude and phase of the  $j$ -th harmonic<sup>1</sup>.

For a particular fundamental frequency  $\omega_0$ , coefficients  $a_k$ ,  $\beta_j$  and  $\alpha_j$  can be solved for by either Least Squares or the Total Least Squers technique [8][12]. The *TLS* solution is discussed below.

For a block of  $N+p$  samples, equations 1 can be written in matrix form:

$$\begin{bmatrix} y_n & \dots & y_{n-p} \\ y_{n-1} & \dots & y_{n-p-1} \\ \dots & & \dots \\ y_{n-N_1} & \dots & y_{n-N_1-p} \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} c_{1,n} & s_{1,n} & \dots & s_{q,n} \\ c_{1,n-1} & s_{1,n-1} & \dots & s_{q,n-1} \\ \vdots & \vdots & \dots & \vdots \\ c_{1,n-N_1} & s_{1,n-N_1} & \dots & s_{q,n-N_1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \alpha_1 \\ \vdots \\ \beta_q \\ \alpha_q \end{bmatrix} =$$

<sup>1</sup> In [11] similar model is used for sinusoids estimation in colored noise.

$$\begin{bmatrix} e_n \\ e_{n-1} \\ \vdots \\ e_{n-N_1} \end{bmatrix} \quad (2)$$

where  $c_{k,i}$ ,  $s_{k,i}$  stand for *cosine* and *sine*  $k$ -th harmonic at time  $i$ . This can be written more compactly as:

$$[y_n, \mathbf{Y}_n, \mathbf{\Omega}][1, \mathbf{a}^t, \gamma^t]^t = [y_n, \mathbf{H}][1, \theta^t]^t = \mathbf{H}^e [1, \theta^t]^t = e_n \quad (3)$$

where  $\mathbf{\Omega}$  is the matrix whose columns consist of  $q$  *sine* and *cosine* waveform harmonics of  $\omega_0$  and  $\gamma$  is a vector of unknown harmonic amplitudes  $\alpha_i, \beta_i; i = 1 : q$  and  $N_1 = N + 1$ . Using *TLS* technique all unknown parameters can be estimated by approximating the  $(p+q+1)$ -rank matrix  $\mathbf{H}^e$  with a rank deficient matrix

$$\hat{\mathbf{H}}^e = [\hat{y}_n, \hat{\mathbf{H}}] = [\hat{y}_n, \hat{\mathbf{Y}}_n, \hat{\mathbf{\Omega}}] \quad (4)$$

i.e. by reconstructing  $\hat{\mathbf{H}}^e$  from SVD of  $\mathbf{H}$  but setting the smallest SV  $\sigma_{p+q+1} = 0$ . The solution is  $\theta = \hat{\mathbf{H}}^{\#} \hat{y}_n$ .

The matrix approximation error is  $\sigma_{p+q+1}$  which is also the intensity of the noise signal  $e_n$ . Varying  $\omega_0$  and looking for a solution that gives minimal  $\sigma_{p+q+1}$ , we get the optimal estimate of model parameters  $\theta$  and the optimal estimate of  $\omega_0$ . The matrix  $\mathbf{\Omega}$  consists of *exact* signals but it is also perturbed by *TLS* technique. That is essential since  $\mathbf{\Omega}$  does not represent the exact excitation but a crude approximation to it.

## 3. DIFFERENTIATING PHASE DATA

The second method that we propose is based on the differentiation of phase data. The approach used here is strongly motivated by the work of Kay [13]. Consider that a speech signal is passed through a filter-bank and that at a particular bank  $k$  the output consists of a single complex sinusoid in complex white Gaussian noise:

$$x(n) = A e^{j(\omega n + \theta)} + z(n), \quad n = 0, 1, 2, \dots, N-1. \quad (5)$$

where  $A$ ,  $\omega$  and  $\theta$  are the amplitude, the frequency and the phase, respectively and the noise  $z(n)$  is assumed to be a zero mean complex white Gaussian process with variance  $\sigma_z^2$ . In case of a high SNR ( $A^2/\sigma_z^2$ ),  $x(n)$  can be approximated by:

$$x(n) \approx A e^{j(\omega n + \theta + u(n))}, \quad n = 0, 1, 2, \dots, N-1. \quad (6)$$

where  $u(n)$  is zero mean white Gaussian noise with variance  $\sigma_z^2/2A^2$ . All the information required to estimate  $\omega$  (and  $\theta$ ) is contained in the phase angle:

$$\phi(n) = \omega n + \theta + u(n) \quad (7)$$

Because we want to estimate only the frequency  $\omega$  we consider the differenced phase data:

$$\Delta(n) = \phi(n+1) - \phi(n) = \omega + u(n+1) - u(n) \quad (8)$$

The problem now is to estimate the mean of a colored Gaussian noise process. In matrix notation, Eq. 8 can be written as:

$$\mathbf{\Delta} = \omega \mathbf{1} + \mathbf{u} \quad (9)$$

where  $\mathbf{\Delta} = [\Delta(0), \Delta(1), \dots, \Delta(N-1)]^T$ ,  $\mathbf{1} = [1, 1, \dots, 1]^T$  and  $\mathbf{u} = [u(1)-u(0), u(2)-u(1), \dots, u(N-1)-u(N-2)]^T$ .

The MLE of  $\omega$  is equivalent to the minimum variance unbiased estimator for the model of Eq. 9 and it can be shown to be:

$$\hat{\omega} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \Delta}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \quad (10)$$

where  $\mathbf{C}$  is the covariance matrix of  $\Delta$ . Note that the model of Eq. 8 is a moving average process with coefficients +1 and -1 with driving noise variance  $\sigma_z^2/2A^2$ . Thus, the inverse of the covariance matrix  $\mathbf{C}$  can be shown [14] to be:

$$c_{ij} = \frac{2A^2}{\sigma_z^2} \left[ \min\{i, j\} - \frac{ij}{N} \right] \quad 1 \leq i, j \leq N-1 \quad (11)$$

Using Eq. 11 we can further show that:

$$\begin{aligned} \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} &= \frac{N(N^2-1)A^2}{6\sigma_z^2} \\ \mathbf{1}^T \mathbf{C}^{-1} \Delta &= \frac{N(N^2-1)A^2}{6\sigma_z^2} \sum_{n=0}^{N-2} w(n)\Delta(n) \end{aligned} \quad (12)$$

where

$$w(n) = \frac{\frac{3}{2}N}{N^2-1} \left\{ 1 - \left[ \frac{n - (N/2 - 1)}{N/2} \right]^2 \right\} \quad (13)$$

Note that:

$$\sum_{n=0}^{N-2} w(n) = 1 \quad (14)$$

Substituting Eq. 12 into Eq. 10 gives the result:

$$\hat{\omega} = \sum_{n=0}^{N-2} w(n)\Delta(n) \quad (15)$$

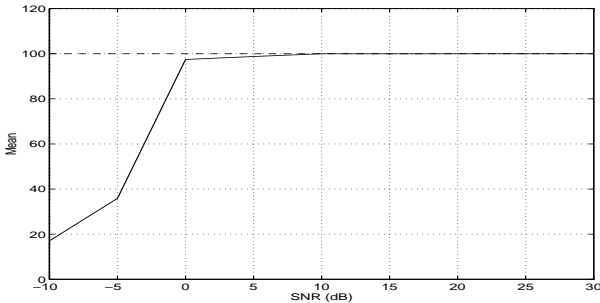
Because Eq. 8 can be written as:

$$\begin{aligned} \Delta(n) &= \text{angle}\{x(n+1)\} - \text{angle}\{x(n)\} \\ &= \text{angle}\{x^*(n)x(n+1)\} \end{aligned} \quad (16)$$

an equivalence to Eq. 15 is:

$$\hat{\omega} = \sum_{n=0}^{N-2} w(n) \text{angle}\{x^*(n)x(n+1)\} \quad (17)$$

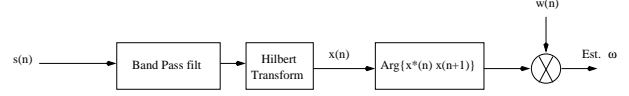
As an example of application of Eq. 17, let's consider a data record of  $N=400$  points of a complex sinusoid with  $\omega = 2\pi \cdot 0.00625$  (sampling freq. 16 kHz, frequency of the sinusoid : 100 Hz). Figure 2 shows the mean square error versus SNR (in decibels) after 100 realizations for each SNR. The above estimation formula can be applied to the



**Figure 2:** Performance of the frequency estimator (Eq. 17).

estimation of the fundamental frequency of speech using

a filter bank or one band-pass (30-600 Hz) filter. In the first case, a linear regression of the estimated frequencies at each band of the filter bank provides an estimation of the fundamental frequency of the input signal. However, we have found that the simple band-pass filter works also well, and this the approach used here. Figure 3 shows the block diagram of the proposed method.



**Figure 3:** Block diagram of the second proposed method.

## 4. RESULTS

Throughout this section, we will refer to the proposed algorithms as ARHE\_PD (ARHE based Pitch Detector) and DP\_PD (Differentiating Phase based Pitch Detector). The performance of the proposed algorithms was evaluated on speech data taken from a database used for speech synthesis (high quality recordings). We used 10 files from a female speaker and 10 files from a male speaker ( $\sim 1$  min.). Pitch values were estimated every 10 ms. Voicing estimates and reference pitch values were obtained using the pitch detector proposed in [5] (RAPT). As a second reference, the pitch detector proposed in [6] was used (SBPD). In contrast to these reference pitch detectors, the proposed methods do not use any smoothing (e.g., median filtering) or constraints. From many previous tests we have found that these two reference algorithms have a similar performance. Because both reference algorithms set the pitch value to zero into unvoiced areas (in contrast to the proposed algorithms where pitch values are also estimated during unvoiced frames), we have evaluated the pitch detection only on voiced frames. The evaluation has been based on the weighted Gross Pitch Error (GPE) [9] criterion, defined as:

$$GPE = \frac{1}{K} \sum_{k=1}^K \left( \frac{E_k}{E_{max}} \right)^{1/2} \left| \frac{\omega_k - \hat{\omega}_k}{\hat{\omega}_k} \right| \quad (18)$$

where  $K$  denotes the number of voiced frames,  $E_k$  the short time energy of the  $k$ th frame:

$$E_k = \sum_{n=0}^{N-1} |x_k(n)|^2 \quad (19)$$

$E_{max}$  represents the maximum short-time energy, and  $\omega_k$  and  $\hat{\omega}_k$  are the reference and estimated pitch frequencies for the  $k$ th frame, respectively. Table 1 shows the performance of the algorithms for clean speech. The reference pitch values were those obtained from RAPT. For comparison purposes, in Tab. 1 we also include the GPE obtained from SBPD. The performance of the proposed algorithms was also evaluated under noisy conditions. A white Gaussian noise of zero mean was added to the clean speech, and the performance was evaluated at SNR's of 10, 5 and 0 dB. Table 2 shows the GPE at different SNR's. Pitch frequency contours (from the DP\_PD method) of a typical male utterance at different SNR's are shown in Fig. 4. It is evident that the algorithm performs satisfactorily in such

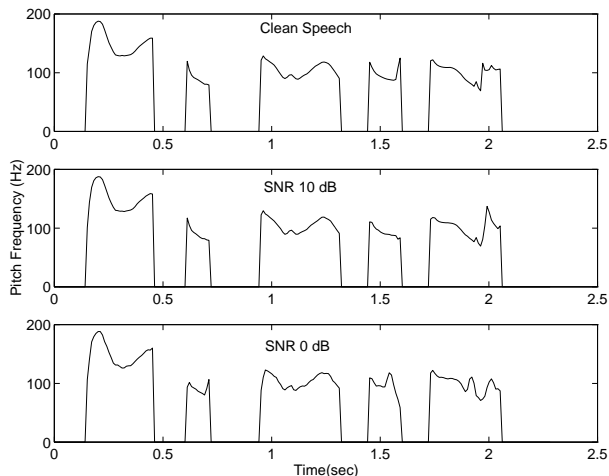
Method	GPE (%)
DP_PD	1.32
ARHE_PD	2.53
SBPD	1.35

**Table 1:** Performance of the proposed algorithms compared to the algorithms RATP and SBPD. The reference pitch values were obtained from RATP.

SNR (dB)	DP_PD	ARHE_PD
10	0.67	1.27
5	0.78	1.36
0	1.20	1.68

**Table 2:** Performance of the proposed algorithms under noise conditions. Reference pitch values are their estimates from the clean data. GPE is in percent.

noise environment. Compared to SBPD, both methods were found to be more robust under these noise conditions (see Tab. 2). The proposed algorithms as well as SBPD and RATP have been used for analysis and copy synthesis of speech using the Harmonic plus Noise Model, HNM [10]. Reconstructed speech was of high quality using all pitch detectors. We found that GPE of more than 3% results in degradations of the reconstructed speech signal.



**Figure 4:** Performance of the proposed algorithm under noise conditions (from DP\_PD method).

## 5. CONCLUSION

In this paper two new methods for the estimation of pitch are proposed. The first method are based on the Autoregressive Harmonic Excitation (ARHE) speech model and second is based on a model of differentiated phase data obtained from the Hilbert transformed speech signal. Phase data are considered to be modeled as a colored Gaussian noise process. The MLE of the fundamental frequency is obtained as the minimum variance unbiased estimator of

the model. The performance of the proposed pitch estimators was evaluated on clean speech as well as under noise conditions. Results shown that the methods have comparable performance to previous reported pitch detectors (e.g., RATP, SBPD) while their performance is maintained under noisy conditions.

## 6. REFERENCES

- [1] W. Hess, *Pitch determination of Speech Signals: Algorithms and Devices*. Berlin: Springer, 1983.
- [2] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Albuquerque), pp. 249–252, 1985.
- [3] S. Seneff, "Real-time harmonic pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 358–365, Aug 1978.
- [4] M. Lahat, R. Niederjohn, and D. Krubsack, "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, pp. 741–750, June 1987.
- [5] D. Talkin, "A robust algorithm for pitch tracking," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), ch. 14, pp. 495–518, Elsevier, 1995.
- [6] Y. Stylianou, "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech," *IEEE Nordic Signal Processing Symposium.*, Sept 1996.
- [7] D. Griffin and J. Lim, "Multiband-excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 236–243, Feb 1988.
- [8] I. Zeljkovic, "Autoregressive harmonic excitation (arhe) speech model and its parameter estimation by total least squares," *Submitted for publication in IEEE Signal Processing Letters*, 1998.
- [9] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech and Audio Processing*, vol. 7(3), pp. 333–338, May 1999.
- [10] Y. Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, Jan 1996.
- [11] C. Chatterjee, R. L. Kashyap, and G. Boray, "Estimation of close sinusoids in colored noise and model discrimination," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-41, pp. 88–101, 1987.
- [12] S. V. Huffel and J. Vandewalle, "The total least squares problem, computational aspects and analysis," *SIAM*, 1991.
- [13] S. Kay, "A fast and accurate single frequency estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1987–1990, Dec 1987.
- [14] B. Noble and J. Daniel, *Applied Linear Algebra*. Englewood Cliffs, New Jersey: Prentice Hall, 1977.