PITCH ESTIMATION USING MUSIC ALGORITHM BASED ON THE SINUSOIDAL SPEECH MODEL

Mehdi Hosseinpour*, Hamidreza Amindavar†

*†Amirkabir University of Technology
*Iranian Telecommunication Research Center
Electrical Engineering Department
Hafez Avenue, 15914
Tehran, Iran
*mhosseinpour@yahoo.com
†hamidami@cic.aku.ac.ir

Abstract– In this paper we present a new method of pitch estimation in speech signals. The existing methods in pitch estimation do not provide an accurate estimate of the pitch value, in presence of sinusoidal noise. Thus these methods have large errors in presence of sinusoidal noise and some other noise characteristics. The method we presented in this paper is based upon the MUSIC algorithm to estimate the pitch frequency. MUSIC algorithm has been widely used to estimate the frequency and angle of arrivals for narrow(wide) band singals. We note that the algorithm presented here accurately estimates the pitch frequency, in the presence of sinusoidal noise and like the other methods in the presence of Gaussian noise.
Index-Terms.— MUSIC, MBE.

1. INTRODUCTION

The error in estimating the pitch frequency of a speech signal occurs under two different groups. The group of errors are known as gross errors, and the second kind of errors are termed as fine errors [1]. The gross pitch error describes the difference between the true and the estimated value is gross, such as halving, and doubling. The fine pitch errors depict how precise is the pitch for the synthesis of high quality speech given the limited amount of resolution available. In multiband excitation (MBE) vocoder [2] the pitch is estimated in fractional resolution by spectral analysis. However, the estimated pitch may result in gross pitch errors. On the other hand, the accuracy in estimating the pitch can have a direct impact on the quality of the synthesized speech. Therefore, in the most MBE vocoders at least eight bits [3] are assigned for the pitch quantization. The improved MBE such as used in the INMARSAT-M standard for satellite voice communication cannot solely provide an accurate estimate of the pitch value. The sole applications of the existing techniques for pitch estimation sometimes result in gross errors. The extraction of the pitch comes significantly more complicated in presence of noise. Although there are some technqiues that estimate the pitch frequency more accurately in presence of Gaussian noise, see [5], but their accuracy is significantly detoriate in sinusoidal noise. The approaches such as MBE or its improved versions, see [6], are not accurate neither in the presence of Gaussian nor in the presence of sinusoidal noise. In order to avoid the major obstacle in MBE vocoders, i.e., lower accuracy in the presence of sinusoidal noise. Therefore, we present our method based on MUSIC approach, this provides a good justification for determining the accurate pitch frequency.

In this paper, we provide an accurate method to estimate pitch frequency based on the MUSIC (Multiple Signal Classificatio) algorithm [4]. This algorithm is an eigen vector based approach to signal processing problems. Under the basic MUSIC method strategy the signal of interest is assumed as the summation of L uncorrelated narrow band signals plus a white noise. According to MUSIC method the center frequency of these narrow band signals are determined.

2. PROBLEM FORMULATION

In this section we discuss the pitch frequency estimation via the MUSIC method. We assume as in all the sinusoidal vocoders model, the speech signal is represented by the following

$$u(n) = \sum_{m=1}^{L} A_m \sin(L\omega_0 n + \theta_m) + v(n), \qquad (1)$$

where $\{A_m\}$, ω_0, and $\{\theta_m\}$ parameters represent the magnitudes, the pitch frequency, and some phases, and $v(n)$ is the common additive noise present in all communication systems, in our treatment $v(n)$ is modelled as a Gaussian or a sinusoidal noise.

Our aim in using the MUSIC method is to determine the pitch frequency ω_0. On the other hand, equation (1) is the suitable model for the MUSIC approach.

We estimate the $(M+1) \times (M+1)$ ensemble-averaged correlation matrix $\mathbf{R}$, where M represents the degree of correlation length, and N is the length of window containing

the frame. By passing $u(n)$ in equation (1) from a low-pass filter of cutoff frequency at 800Hz to avoid the possible disturbance by the undesireable frequencies beyond the 500Hz. Since the speech signal is realistically noisy we assume that there are L different sinusoids with angular frequencies at $\{\omega_1, \omega_2, \cdots, \omega_L\}$ and their respective powers, $\{P_1, P_2, \cdots, P_L\}$. For implementing the MUSIC we are required to estimate the ensemble-averaged autocorrelation matrix $\mathbf{R}$, without loss of generality we assume this matrix is estimated by the following

$$\hat{\mathbf{R}} = \frac{1}{2(N-M)}\Phi, \quad \Phi = A^T A, \tag{2}$$

where Φ is related to the data matrix A. The data matrix is given by

$$A^T = \begin{bmatrix} u(M) & u(M+1) & \cdots & u(N) \\ u(M-1) & u(M) & \cdots & u(N-1) \\ \vdots & \vdots & \vdots & \vdots \\ u(1) & u(2) & \cdots & u(N-M+1) \end{bmatrix} \tag{3}$$

let the $\{v_1, v_2, \cdots, v_{M+1}\}$ denote the eigenvectors of the estimated autocorrelation matrix. We define the eigenvectors $\{v_1, v_2, \cdots, v_L\}$ associated with the L largest eigevalues of $\hat{\mathbf{R}}$, and $\{v_{L+1}, v_{L+2}, \cdots, v_{M+1}\}$ corresponding to the $(M+1-L)$ smallest eigenvalues of $\hat{\mathbf{R}}$. If we denote

$$V_N = [v_{L+1}, v_{L+2}, \cdots, v_{M+1}], \tag{4}$$

the power spectrum of the $u(n)$ based on the MUSIC method is determined by

$$\begin{aligned} S_{\text{music}}(\omega) &= \frac{1}{\sum_{i=L+1}^{M+1} |\mathrm{s}^H v_i|^2} \\ &= \frac{1}{\mathrm{s}^H V_N V_N^T \mathrm{s}}, \end{aligned} \tag{5}$$

where

$$\mathrm{s}^T = \left[1, e^{-j\omega}, \cdots, e^{-j\omega(M-L)}\right], \quad |\omega| \leqslant \pi.$$

In order to estimate the pitch frequency based on the MUSIC, we first form the data matrix A where the speech signal is sampled 8000Hz, and the window length is 256, i.e.,N=256, and for simulation purposes we assume M=16. The actual value of M can also be estimated via the minimum description length method or the Akaike information criterion. We subsequently use the singular value decomposition approach to determine the smallest singular values $\{\sigma_{L+1}, \sigma_{L+2}, \cdots, \sigma_{M+1}\}$ and the corresponding singular vectors $\{v_{L+1}, v_{L+2}, \cdots, v_{M+1}\}$ in equation (4). We determine the approximate spectrum of the speech singal in the window length of N using equation (5). The more the amplitude of the peak at a specific frequency the higher is the likelihood of the presence of a sinusoidal signal with such a frequency.

3. PITCH DETECTION

The presence of $v(n)$ in (1) can result in erroneous peaks far from the pitch frequency, we therefore present a strategy to detect the correct pitch in the following. Let $\{C_1, C_2, \cdots, C_n\}$ are the estimates of the peaks obtained from MUSIC algorithm. Since the larger peaks in MUSIC algorithm are to denote the presence of strong sinusoidal signal we determine the average of the peaks, $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$, and group the peaks higher than $\bar{C}$, hence, we have determined a group of candidate frequencies as the correct harmonic frequencies. Let $\{C_1', C_2, \cdots, C_k'\}$ are the selected peaks, and the corresponding angular frequencies are $\{\omega_1', \omega_2', \cdots, \omega_k'\}$, and $\omega_h = \min\left\{\omega_i' \Big|_{i=1}^{k}\right\}$ is taken as the first estimate of the pitch frequency. This is a realistic model because in speech signal the harmonics are not an integer multiple of a single frequency. We therefore can accept the following model
$omega_0 = \omega_h + d$, and we assume an ω_0. For the selected ω_0 for each peak the closest angular frequency from $\{\omega_1', \omega_2', \cdots, \omega_k'\}$ which is an integer multiple of ω_0 is taken as the estimate of the harmonic frequency. Let $\{\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_k\}$ be the set of the estimates of the harmonics, we then obtain the error $\{\hat{\omega}_1 - \omega_1', \hat{\omega}_2 - \omega_2', \cdots, \hat{\omega}_k - \omega_k'\}$, we subsequently select the ω_0 with smallest residual error. In figure 1 we show a speech signal for a female, and figure 2 and the error from the MBE is shown where the estimated pitch frequency is 214.310Hz. In figure 3 we present the result for MUSIC method where the pitch frequency according to this introduced method is estimated as 215.75Hz, which is the exact value.

4. SIMULATION

In order to verify the introduced method based on the MUSIC approach we conducted the following experiments. The MUSIC and MBE methods for a several frames of a clean speech are tested. To determine the error in each pitch estimation, the "true" pitch was manually estimated from observations of the waveform and the spectrum of the residual using a graphical tool. Then the normalized pitch error was computed by dividing the pitch error by the pitch value and was expressed in percentages. We consider the normalized pitch error in percentage as a low gross error if it is between 10% to 30% and as a high gross error if it is higer than 30%. In figures 4, 5, and 6 we show the pitch contour for a male clean speech using the actual, MUSIC, and MBE method. In table 1 we show the results of MUSIC and MBE for a clean speech. The MUSIC and MBE methods in presence of Gaussian noise with SNR=0, -10dB, are examined and the results are presented in tables 2 and 3. In table 4 and 5, the results of MUSIC and MBE in presence of sinusoidal noise with SNR=0, -10dB are shown. We note that MUSIC outperforms the MBE approach.

5. CONCLUSION

In this paper a new approach to determine the pitch frequency is described. This method is based on the MUSIC approach where the speech window is modelled as a summation of L sinusoides. The eigen based MUSIC method is used to extract the narrow band signals present in the speech. These narrow band signals have their center frequency as the pitch frequency. A key point in using this new approach is its independence from the quality of the

SNR=∞	Male		Female		Overall
	Gross errors		Gross errors		
	High %	Low %	High (%)	Low (%)	Gross error (%)
MUSIC	0.07	0.36	0.89	0.21	0.38
MBE	0.64	0.35	0.46	0.32	0.44

Table 1: Clean speech gross errors with different methods.

SNR=0	Male		Female		Overall
	Gross errors		Gross errors		
	High %	Low %	High (%)	Low (%)	Gross error (%)
MUSIC	2.73	2.04	5.72	0.12	2.65
MBE	6.15	4.09	3.35	1.34	3.73

Table 2: Noisy speech gross errors with different methods, Gaussian noise.

origin of the signal, i.e., noisy and clean speech. We compared the performance of the MUSIC approach against the method based on the normalized autocorrelation and the MBE method, where we showed through some simulations that MUSIC can have superior performance.

6. REFERENCES

[1] W. J. Hess, "Pitch and voicing determination," *Advances in speech signal processing*, ed., by S. Furui and M. M. Sondhi, Marcel Dekker Inc., pp.3-48, 1992.

[2] W. Griffin, J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. on ASSP*, Vol. 36, pp. 1223-1235, August 1988.

[3] S. Yeldener, A. M. Kondoz, B. G. Evans, "High quality multiband LPC coding of speech at 2.4kbit/s," *Electronic letters*, Vol. 27, No. 14, July 1991.

[4] S. Haykin, "Adaptive filter theory," 3rd ed., Prentice Hall, 1996.

[5] D. A. Krubsack, R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with coherence measures developed for noise-corrupted speech," *IEEE Trans. on Signal processing*, Vol. 39, No. 2, pp. 319-329, Feb. 1991.

[6] C. -F. Chan, E. W. M. Yu, "Improving pitch estimation for efficient multiband excitation coding of speech," *Electronic letters*, Vol. 32, No. 10, May 1996.

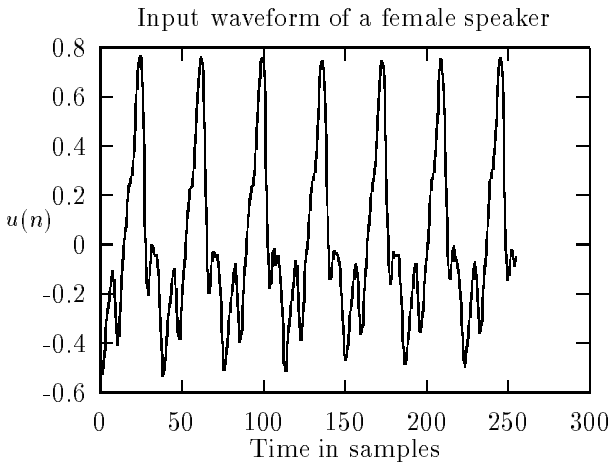

Figure 1: Input waveform of a female speaker sampled at 8000Hz.

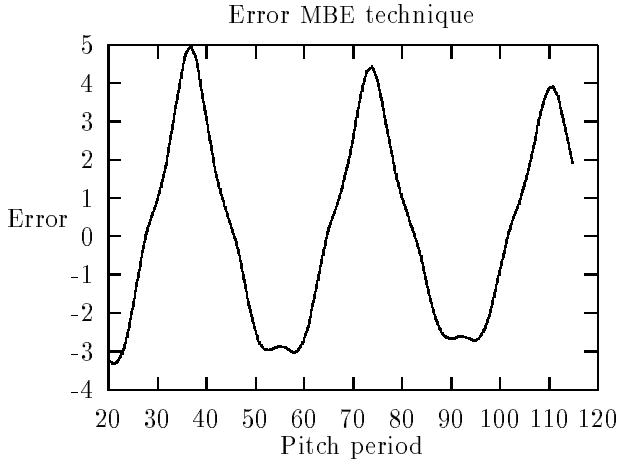

Figure 2: The error performance of the MBE method.

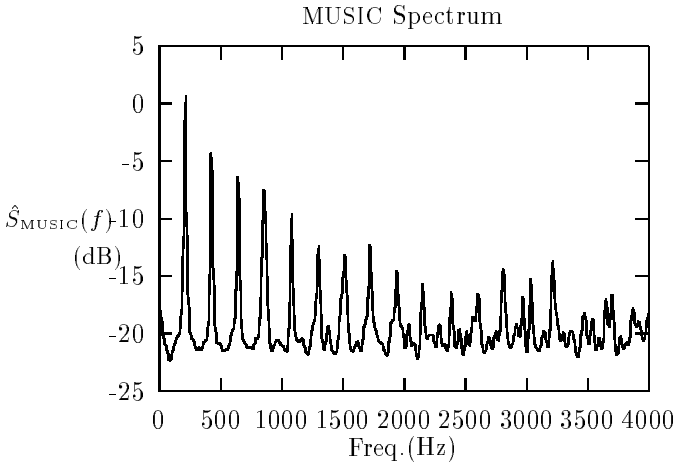

Figure 3: The spectrum of MUSIC method for a female speaker.

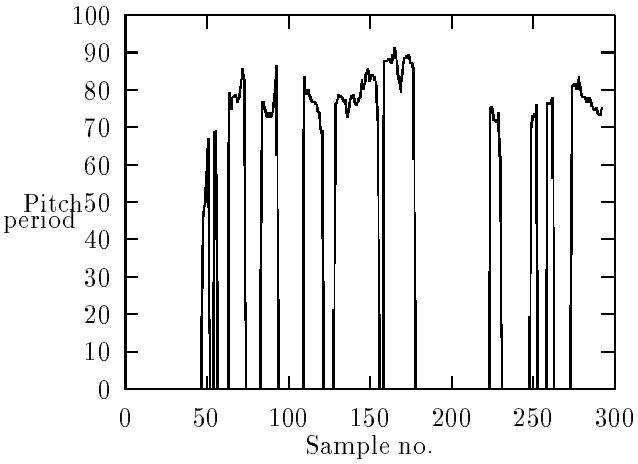

Figure 4: The typical actual pitch contour curve.

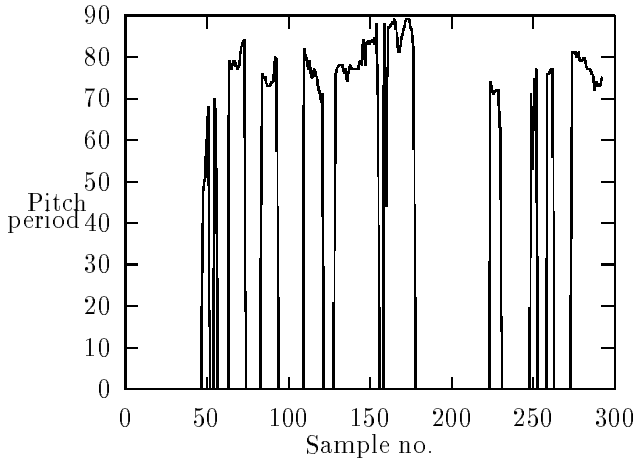

Figure 5: The typical pitch contour curve for the MBE method.

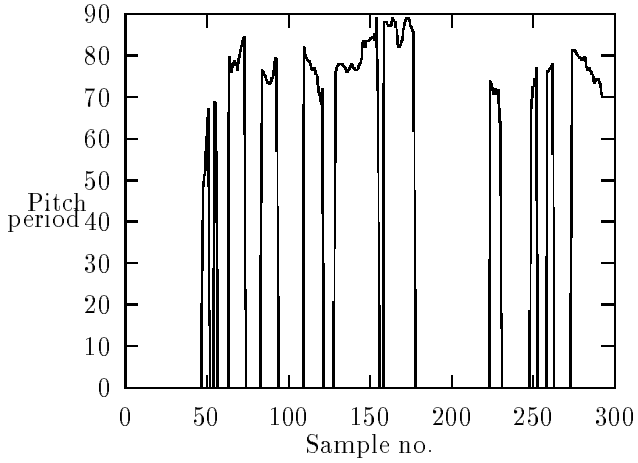

Figure 6: The typical pitch contour curve for the MUSICmethod.

SNR=-10	Male		Female		Overall
	Gross errors		Gross errors		
	High %	Low %	High (%)	Low (%)	Gross error (%)
MUSIC	15.01	10.23	14.13	4.69	11.02
MBE	14.67	8.53	16.14	12.44	12.94

Table 3: Noisy speech gross errors with different methods, Gaussian noise.

SNR=0	Male		Female		Overall
	Gross errors		Gross errors		
	High %	Low %	High (%)	Low (%)	Gross error (%)
MUSIC	0.34	1.71	1.34	2.12	1.37
MBE	18.08	1.36	22.93	3.17	11.39

Table 4: Noisy speech gross errors with different methods, Sinusoidal noise.

SNR=-10	Male		Female		Overall
	Gross errors		Gross errors		
	High %	Low %	High (%)	Low (%)	Gross error (%)
MUSIC	1.64	2.73	1.56	2.68	2.15
MBE	36.17	3.41	32.24	12.85	21.16

Table 5: Noisy speech gross errors with different methods, Sinusoidal noise.