Joint Time Delay and Pitch Estimation for Speaker Localization

L.Y. Ngan[1], Y. Wu[2], H.C. So[3], P.C. Ching[1] and S.W. Lee[1]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong
[2]Laboratory for Radar Signal Processing, Xidian University
[3]Department of Computer Engineering and Information Technology, City University of Hong Kong
lyngan@ee.cuhk.edu.hk, wuyuntao6@163.com, ithcso@cityu.edu.hk, pcching@ee.cuhk.edu.hk and yswlee@ee.cuhk.edu.hk

ABSTRACT

In this paper, we attempt to develop an efficient and accurate algorithm for joint time delay and pitch estimation of a speech signal received at a microphone array. The time delay measurement allows a speaker to be located while the detection of the pitch frequency is useful for analyzing the acoustic properties of the sound. A subspace method based on state-space realization is first introduced for joint time delay and frequency estimation of a synthetic signal consisting of several frequency components. The frequency estimates are obtained directly from the eigenvalues of the state transition matrix whilst the time delay is approximated from the observation matrix and the estimated frequencies using a least square approach. The method is then extended to track both the time delay and pitch frequency of a speech signal, which is modeled by a summation of sinusoids that are harmonically related to the fundamental frequency (pitch) and spectrally shaped by the vocal tract transfer function. Extensive simulation tests have been done to validate the effectiveness and accuracy of the proposed algorithm.

1. INTRODUCTION

In applications such as hands-free communications and videoconferencing, it is necessary to know accurately the position of the speaker for microphone and camera steering. While on the other hand, for recognition purpose, in particular for tonal languages like Chinese and other Asian languages, it is also required to keep track the pitch profile of the spoken utterances. Localization is usually achieved by triangulation using time-difference-of-arrival (TDOA) measurements and there are many existing localization algorithms in the literature [1]-[2]. Accurate extraction of the fundamental frequency of a voiced sound is by no mean trivial although there exist many different methods with various levels of complexity [3]-[4].

Recently, the problem of joint time delay and frequency estimation of sinusoidal signals has attracted considerable attention. Applications include speech enhancement and pitch estimation [5]-[6], synchronization in CDMA systems [7] and FSK demodulation using multiple segments [8].

Let the received sensor outputs be

$$\begin{cases} r_1(n) = s(n) + w_1(n) \\ r_2(n) = s(n-\tau) + w_2(n) \end{cases}, \quad n = 0,1,\cdots,N-1 \qquad (1)$$

where $$s(n) = \sum_{m=1}^{P} \alpha_m \exp(j\omega_m n) \qquad (2)$$

The source signal $s(n)$ is represented by a sum of P complex sinusoids of which the amplitudes $\{\alpha_m\}$ are unknown constants and the normalized radian frequencies $\{\omega_m\}$ are distinct. Without loss of generality, we assume $|\alpha_1| > |\alpha_2| > \ldots > |\alpha_P|$. The number of sinusoids present in $s(n)$, P, is assumed to be known *a prior*. The additive noises, $w_1(n)$ and $w_2(n)$, are both uncorrelated zero-mean complex white Gaussian processes with the same variance σ_w^2. The parameter τ denotes the difference in arrival times at the two receivers and N is the number of samples collected at each channel. The objective is to estimate both the time difference of the received signals and the frequencies of their constituent components.

Sherman *et al* [9] developed an ESPRIT algorithm to estimate τ when $P > 1$ while a generalized Yule-Walker solution was suggested in [10] to determine $\{\omega_m\}$ separately. Qian and Kumaresan [5] proposed a subspace-based method for joint time delay and frequency estimation where the estimates are obtained using the eigenvalues and eigenvectors of a matrix derived from the covariance matrices of the received signals. This technique derives the frequencies from the eigenvectors, and the accuracy of the estimation is thus limited. On the other hand, a state-space model for multiple sinusoidal frequency estimation has been reported in [11], which was motivated by the fact that state-space parameterization enables reduction of parameter sensitivity [12]. This model had also been employed to find the direction-of-arrival (DOA) of an object [13] as well as to facilitate joint DOA and frequency estimation [14] using array data.

In this paper, we shall introduce a new subspace approach for estimating the frequency components of a composite sinusoidal signal received at two spatially separated sensors as well as their differential delay. Unlike the method reported in [6] where the joint delay and frequency estimation is achieved by using the eigenvectors of a matrix derived from the covariance matrix of the received signals, the proposed method here is based on a state-space realization, and the frequencies and time delay are obtained from the transition and observation matrices instead. The performance of the proposed estimator is compared with the Cramér-Rao lower bound (CRLB) and it is found that the result is better than that of [6]. The method is then extended to track both the time delay and pitch frequency of a stream of real speech samples arriving at two separated microphone sensors.

2. FORMULATION OF PROPOSED METHOD

Using (1), we first form the following sets of signal vectors:

0-7803-7761-3/03/$17.00 ©2003 IEEE

$$\mathbf{X}_1(k) = [r_1(k), r_1(k+1), \cdots, r_1(k+M-1)]^T$$
$$\mathbf{X}_2(k) = [r_2(k), r_2(k+1), \cdots, r_2(k+M-1)]^T \quad (3)$$

where $k = 0, 1, \ldots, K-1$ and $K = N - M + 1$ and T denotes the transpose operation. The parameter M is the length of each vector and its value lies between $P + 1$ and $N - P + 1$ so that the span of any K of $\mathbf{X}_1(k)$ and $\mathbf{X}_2(k)$ has no rank less than P. Substituting (1) and (2) into (3) yields

$$\mathbf{X}_1(k) = \mathbf{A}(\omega)\mathbf{S}(k) + \mathbf{W}_1(k)$$
$$\mathbf{X}_2(k) = \mathbf{A}(\omega)\mathbf{\Delta}(\omega,\tau)\mathbf{S}(k) + \mathbf{W}_2(k) \quad (4)$$

where
$$\mathbf{A}(\omega) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{j\omega_1} & e^{j\omega_2} & \cdots & e^{j\omega_P} \\ \vdots & \vdots & \vdots & \vdots \\ e^{j\omega_1(M-1)} & e^{j\omega_2(M-1)} & \cdots & e^{j\omega_P(M-1)} \end{bmatrix} \quad (5)$$

$$\mathbf{S}(k) = [\alpha_1 e^{j\omega_1 k}, \alpha_2 e^{j\omega_2 k}, \cdots, \alpha_P e^{j\omega_P k}]^T$$
$$\mathbf{\Delta}(\omega,\tau) = diag(\alpha_1 e^{j\omega_1 \tau}, \alpha_2 e^{j\omega_2 \tau}, \cdots, \alpha_P e^{j\omega_P \tau})$$
$$\mathbf{W}_i(k) = [w_i(k), w_i(k+1), \cdots, w_i(k+M-1)]^T, \quad i = 1,2$$

Using $\mathbf{S}(k)$ as the state vector, and grouping $\mathbf{X}_1(k)$ and $\mathbf{X}_2(k)$ into one vector, namely, $\mathbf{X}(k) = [\mathbf{X}_1^T(k)\ \mathbf{X}_2^T(k)]^T$, we can formulate a state-space model for the array output as

$$\mathbf{S}(k+1) = \mathbf{\Phi}\mathbf{S}(k) \quad \text{and} \quad \mathbf{X}(k) = \mathbf{B}\mathbf{S}(k) + \mathbf{W}(k) \quad (6)$$

where
$$\mathbf{\Phi} = diag(e^{j\omega_1}, e^{j\omega_2}, \cdots, e^{j\omega_P})$$
$$\mathbf{B} = \begin{bmatrix} \mathbf{A}(\omega) \\ \mathbf{A}(\omega)\mathbf{\Delta}(\omega,\tau) \end{bmatrix} \quad \text{and} \quad \mathbf{W}(k) = \begin{bmatrix} \mathbf{W}_1(k) \\ \mathbf{W}_2(k) \end{bmatrix} \quad (7)$$

Here, $\mathbf{\Phi}$ and $\mathbf{B}$ are the state transition matrix and observation matrix respectively. The estimation algorithm is based on the canonical variables [13], which employs the structure of the cross correlation between future and past data. Let $L \geq 2$ be an integer and define a $2ML$-vector of the form

$$\mathbf{X}_L(k) = [\mathbf{X}^T(k), \mathbf{X}^T(k+1), \cdots, \mathbf{X}^T(k+L-1)]^T \quad (8)$$

By iterating the state-space model, we can easily get

$$\mathbf{X}_L(k) = \mathbf{\Omega}_L\mathbf{S}(k) + \mathbf{E}_L(k), \quad k = 1,2,\cdots,K-L+1 \quad (9)$$

where
$$\mathbf{\Omega}_L = [\mathbf{B}^T\ (\mathbf{B}\mathbf{\Phi})^T \cdots (\mathbf{B}\mathbf{\Phi}^{L-1})^T]^T \quad (10)$$

is another observation matrix of dimension $2ML \times P$ and $\mathbf{E}_L(k)$ is defined conformably with $\mathbf{X}_L(k)$. It is required that $\mathbf{\Omega}_{L-1}$ must have full column-rank of P and a necessary condition for it is $2M(L-1) \geq P$. If $P \leq 2M$, a sufficient condition is that $\mathbf{B}$ has full column rank.

The cross covariance matrix between the future and past outputs, namely, $\mathbf{X}_L(k)$ and $\mathbf{X}_L(k-L)$, can be easily shown to be

$$\mathbf{R}_{XL}(L) = E\{\mathbf{X}_L(k)\mathbf{X}_L^H(k-L)\} = \mathbf{\Omega}_L\mathbf{R}_S\mathbf{\Omega}_L^H \quad (11)$$

where $\mathbf{R}_S = \text{diag}(|\alpha_1|^2 e^{jL\omega_1}, |\alpha_2|^2 e^{jL\omega_2}, \ldots, |\alpha_p|^2 e^{jL\omega_p})$. Hence, $\mathbf{R}_{XL}(L)$ has low rank equal to P, and its column space coincides with that of $\mathbf{\Omega}_L$. A set of orthonormal basis vectors for the column space can be computed from the P principal left singular vectors of $\mathbf{R}_{XL}(L)$. In practice, we apply singular value decomposition (SVD) to the sample cross covariance $\hat{\mathbf{R}}_{XL}(L)$ which is of the form:

$$\hat{\mathbf{R}}_{XL}(L) = \sum_{k=L}^{K-L+1} \mathbf{X}_L(k)\mathbf{X}_L^H(k-L)/K = \sum_{l=1}^{2ML} \gamma_l \mathbf{u}_l \mathbf{v}_l^H \quad (12)$$

where $\{\gamma_l\}$ represent the singular values arranged in decreasing order while $\mathbf{u}_l$ and $\mathbf{v}_l$ denote the left and right singular vectors respectively. When N and/or the signal-to-noise ratio (SNR) is sufficiently large, say, $N \cdot SNR > 100$, we have

$$span(\hat{\mathbf{\Omega}}) \approx span(\mathbf{\Omega}_L) \quad (13)$$

where $\hat{\mathbf{\Omega}} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_P]$. By asymptotic analysis, $\hat{\mathbf{\Omega}}$ is consistent in the sense that

$$\mathbf{\Omega} = \lim_{N\to\infty} \hat{\mathbf{\Omega}} = \mathbf{\Omega}_L\mathbf{T} \quad (14)$$

for some unknown full rank P x P state transformation matrix $\mathbf{T}$. From $\hat{\mathbf{\Omega}}$, we can find the corresponding transformed state matrices for $\mathbf{\Phi}$ and $\mathbf{B}$:

$$\mathbf{\Phi}_T = \mathbf{T}^{-1}\mathbf{\Phi}\mathbf{T} \quad \text{and} \quad \mathbf{B}_T = \mathbf{B}\mathbf{T} \quad (15)$$

and they are estimated by

$$\hat{\mathbf{B}}_T = \hat{\mathbf{\Omega}}_{1:1} \quad \text{and} \quad \hat{\mathbf{\Phi}}_T = \hat{\mathbf{\Omega}}_{1:L-1}^{\#}\hat{\mathbf{\Omega}}_{2:L} \quad (16)$$

where $\mathbf{\Omega}_{k:l}$ denotes the $2M$ x L block rows from k through l and # is the pseudo-inverse operation. Note that

$$\mathbf{\Omega}_T = \lim_{N\to\infty} \hat{\mathbf{\Omega}}_T \quad \text{and} \quad \mathbf{B}_T = \lim_{N\to\infty} \hat{\mathbf{B}}_T \quad (17)$$

and (16) corresponds to a least square solution for $\hat{\mathbf{\Omega}}_{1:L-1}\hat{\mathbf{\Phi}}_T \approx \hat{\mathbf{\Omega}}_{2:L}$.

Since $\mathbf{\Phi}$ is a diagonal matrix with eigenvalues $e^{j\omega_m}$, $m = 1, 2, \ldots, P$, estimation of $\mathbf{\Phi}$ is achieved via diagonalizing $\hat{\mathbf{\Phi}}_T$ as :

$$\hat{\mathbf{\Phi}}_T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} \quad (18)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_P)$ and $\lambda_1 > \lambda_2 > \ldots > \lambda_P$. From (16) to (18), estimates of $\mathbf{\Phi}$ and $\mathbf{T}$ are given by $\mathbf{\Lambda}$ and $\mathbf{U}^{-1}$ respectively. Therefore, frequency estimates can be computed from [15]

$$\hat{\omega}_m = \angle\lambda_m, \quad m = 1, 2, \cdots, P \quad (19)$$

In [6], the frequency components were calculated from the eigenvectors of $\mathbf{R}_{XL}$ and the performance was somewhat limited. Whereas in this new method, we compute the frequency estimates from the eigenvalue phases of the state transition matrix $\mathbf{\Phi}$ which gives better estimation results.

On the other hand, $\mathbf{B}$ is estimated as $\hat{\mathbf{B}} = \hat{\mathbf{B}}_T\mathbf{U}$. Since the upper and lower parts of $\mathbf{B}$ are $\mathbf{A}(\omega)$ and $\mathbf{A}(\omega)\mathbf{\Delta}(\omega, \tau)$ respectively, we have $\hat{\mathbf{B}}_{1:M}\hat{\mathbf{\Delta}}(\omega, \tau) \approx \hat{\mathbf{B}}_{M+1:2M}$ where $\hat{\mathbf{\Delta}}(\omega, \tau)$ represents the estimate of $\mathbf{\Delta}(\omega,\tau)$. A least square solution for $\hat{\mathbf{\Delta}}(\omega, \tau)$ is $\hat{\mathbf{\Delta}}(\omega, \tau) \approx \hat{\mathbf{B}}_{1:M}^{\#}\hat{\mathbf{B}}_{M+1:2M}$, and the time delay can be approximated using all diagonal elements of $\hat{\mathbf{\Delta}}(\omega, \tau)$ and $\{\hat{\omega}_m\}$:

$$\hat{\tau} = \frac{\sum_{m=1}^{P} \angle\hat{\mathbf{\Delta}}_{m,m}(\omega,\tau)\hat{\omega}_m}{-\sum_{m=1}^{P} \hat{\omega}_m^2} \quad (20)$$

Equation (20) gives a time delay estimate using a least square approach which can remove the bias of the estimator thereby

giving a more accurate time difference between the received signals.

3. PITCH ESTIMATION

In section 2, equation (19) extracts the frequency components of $s(n)$ from the eigenvalues of the state transition matrix, and permutates in the order according to their magnitudes. If the signal $s(n)$ is a voiced sound, then based on the speech production model [16], we can assume that it is obtained by exciting an FIR filter with a periodic impulse train. The period of the impulse response is the pitch whilst the frequency response of the filter is governed by the vocal tract transfer function. Therefore, $s(n)$ can be expressed as

$$s(n) = \sum_{m=1}^{P} \alpha_m \exp(jq_m(2\pi f0)n) \qquad (21)$$

where $q_m \in [1, P]$ and is a distinct integer. In this case, all the frequency components are harmonically related to the fundamental frequency $f0$, and plotting α_m versus frequency will show the formant structure. The number of harmonics present will depend on the bandwidth of the signal. If we limit the frequency band to cover one formant only, then the fundamental frequency or the pitch can be extracted by subtracting the frequencies of two adjacent $\angle\lambda_m$ given by (19). Otherwise, we can first low-pass filter the signal and then determine P based on the number of principal pairs of singular values $\{\gamma_l\}$ present in the cross covariance matrix $\mathbf{R}_{XL}(L)$. By estimating the first P frequencies using (19), $f0$ can be obtained from

$$\hat{f0} = \frac{\sum_{m=1}^{P} \hat{\omega}_m}{2\pi \sum_{m=1}^{P} q_m} = \frac{\sum_{m=1}^{P} \hat{\omega}_m}{\pi(P+1)P} \qquad (22)$$

This estimation can be verified by checking the frequency difference between adjacent pairs of components given by $\angle\lambda_m$.

4. SIMULATION RESULTS

Computer simulations were conducted to evaluate the joint time delay and frequency estimation performance of the proposed method in the presence of white Gaussian noise. All results provided were averages of 400 independent runs for different SNRs. The delayed signal is generated by the method proposed in [17] and different SNRs are obtained by proper scaling of the noise sequences. The block size M is selected to be $2f_s / f_{min}$, where f_s is the sampling frequency and f_{min} is the minimum frequency we can detect, which is selected to be 10kHz and 50Hz respectively in the following simulations.

4.1 Complex Sinusoids

In the first simulation, the source signal $s(n)$ is a sinusoidal signal of the form $s(n) = \alpha_1 e^{j\omega_1 n} + \alpha_2 e^{j\omega_2 n}$ with $\alpha_1 = \alpha_2 = 1/\sqrt{2}$, $\omega_1 = 0.04\pi$ rad/s and $\omega_2 = 0.08\pi$ rad/s. The time delay τ is selected to be $6.4T_s$, where T_s is the sampling period. The mean-square-error (MSE) of frequency estimation are shown in Figure 1(a). We found that the MSE of the proposed method is close to the CRLB, with a difference of less than 1dB when SNR $\geq$ -10dB. In addition, it outperforms the ESPRIT method by about 10dB. For the time delay estimation performance shown in Figure 1(b), the proposed method has a MSE close to the CRLB with a difference of about 0.5dB. The performance for both time delay and frequency estimation is better than that in [6].

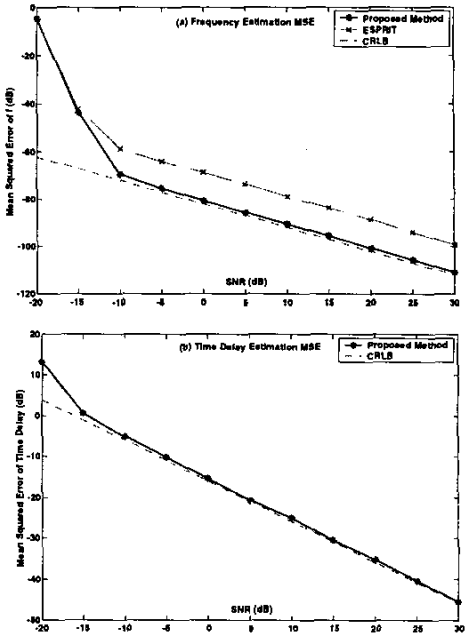

Figure 1: Complex Sinusoids

4.2 Synthetic Speech

In the second simulation, the application of joint time delay and pitch estimation of voiced speech received at two microphones is investigated. A synthetic vowel "EY" is generated by a speech production model with $f0$ equals to 140Hz and the formant frequencies occur at 480Hz, 1720Hz and 2520Hz. The generated speech waveform is shown in Figure 2(a). The time difference of the two microphones is set to be $5.2T_s$. In Figure 2(b), we found that the time delay estimation performance is satisfactory when SNR $\geq$ -10dB. The estimated pitch is found to be 136Hz, which is fairly close to the actual fundamental frequency $f0$.

4.3 Real Speech

In the last simulation, the performance of the proposed method in real speech is investigated. A vowel of about 0.2s is extracted from a continuous Cantonese sentence and the utterance is low-pass filtered at 500Hz. The time delay is selected to be $6.5T_s$. The speech waveform and the time delay MSE with various SNRs are shown in Figure 3. The time difference and frequency components of the speech signal are estimated by (19) and (20). There are only two principal pairs of singular values in $\mathbf{R}_{XL}(L)$, then we can say only two harmonics are present, and the pitch frequency is thus calculated by (22). We found that the joint time delay and pitch estimation performance of the proposed method for real speech is as good as that for the synthetic speech. The estimated pitch frequency is about 229Hz, which is about the same as that calculated by using the Average Magnitude Difference Function (AMDF).

5. CONCLUSIONS

A subspace algorithm based on state-space realization is proposed for joint time delay and pitch estimation of speech signals received at two microphones. The frequencies of the constituent components are obtained directly from the eigenvalues of the state transition matrix whilst the time delay is determined using the observation matrix and the estimated frequencies. The pitch frequency *f0* of a voiced segment can be calculated from the set of estimated frequencies which are harmonically related. It is shown that the time delay estimation performance is satisfactory for both synthetic and real speeches. The pitch frequency *f0* can also be estimated accurately, which allows the pitch profile of a spoken utterance to be traced. By making use of both the time delay and pitch information, one can also track the position of a particular speaker by using a microphone array.

6. REFERENCES

[1] "Special Issue on Time Delay Estimation", *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. 29, pp. 582-5887, June 1981

[2] A.H. Quazi, "An Overview on the Time Delay Estimate in Active and Passive Systems for Target Localization", *IEEE Trans. Acoust. Speech Signal Processing*, Vol. 29, no. 3, pp. 527-533, 1981

[3] K.K. Paliwal and A. Aarskey, "A Comparative Performance Evaluation of Pitch Estimation Methods for TDHS/Sub-band Coding of Speech", *Speech Communication*, Vol. 3, pp. 253-259, 1984

[4] P. Stoica, "List of References on Spectral Line Analysis", *Signal Processing*, Vol. 31, no. 3, pp. 329-340, April 1993

[5] X. Qian and R. Kumaresan, "Joint Estimation of Time Delay and Pitch of Voiced Speech Signals", *Conf. Rec. of the 29th Asilomar Conf. Signals, Systems & Computers*, Vol. 1, pp. 735-739, Nov. 1995, Pacific Grove, CA, USA

[6] G. Liao, H.C. So and P.C. Ching, "Joint Time Delay and Frequency Estimation of Multiple Sinusoids", *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 3121-3124, May 2001, Salt Lake City, Utah, USA

[7] S.R. Dooley and A.K. Nandi, "Adaptive Time Delay and Frequency Estimation for Digital Signal Synchronization in CDMA Systems", *Conf. Rec. of the 32nd Asilomar Conf. Signals, Systems & Computers*, Vol. 2, pp. 1838-1842, Nov. 1998, Pacific Grove, CA, USA

[8] J.A. Sills and Q.R. Black, "Frequency Estimation from Short Pulses of Sinusoidal Signals", *Proc. IEEE MILCOM '96*, Vol. 3, pp. 979-983, 1996

[9] D.L. Sherman, Y.C. Tsai, L.A. Rossell, M.A. Mirski and N.V. Thakor, "Narrowband Delay Estimation for Thalamocortical Epileptic Seizure Pathways", *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Vol. 5, pp. 2939-2942, May 1995, Detroit, Michigan, USA

[10] S.M. Kay, Modern Spectral Estimation: Theory and Applications, Englewood Cliffs, NJ, Prentice Hall, 1988

[11] S.Y. Kung, K.S. Arun and B.D. Rao, "State Space and SVD based Approximation Methods for the Harmonic Retrieval Problems", *J. opt. Soc. Amer.*, Vol. 73, pp. 1799-1811, Dec. 1983

[12] B.D. Rao, "Sensitivity Considerations in State-space Model-based Harmonic Retrieval Methods", *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. 37, No. 11, pp. 1789-1794, Nov. 1989

[13] S. Prasad and B. Chandna, "Direction of Arrival Estimation using Stochastic Model Order Reduction via State Space Modeling", Signal Processing, Vol. 23, No. 2, pp. 157-177, May 1991

[14] M. Viberg and P. Stoica, "A computationally Efficient Method for Joint Direction Finding and Frequency Estimation in Colored Noise", *Conf. Rec. of the 32nd Asilomar Conf. Signals, Systems & Computers*, Vol. 2, pp. 735-739, Nov. 1998, Pacific Grove, CA, USA

[15] Y. Wu, H.C. So and P.C. Ching, "Joint Time Delay and Frequency Estimation via State-Space Realization", submitted to *IEEE Signal Processing Letters*

[16] J.R. Deller, Jr., J.H.L Hansen and J.G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press 2000

[17] P.C. Ching and Y.T. Chan, "Adaptive Time Delay Estimation with Constraints", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 36, pp. 599 –602, April 1988

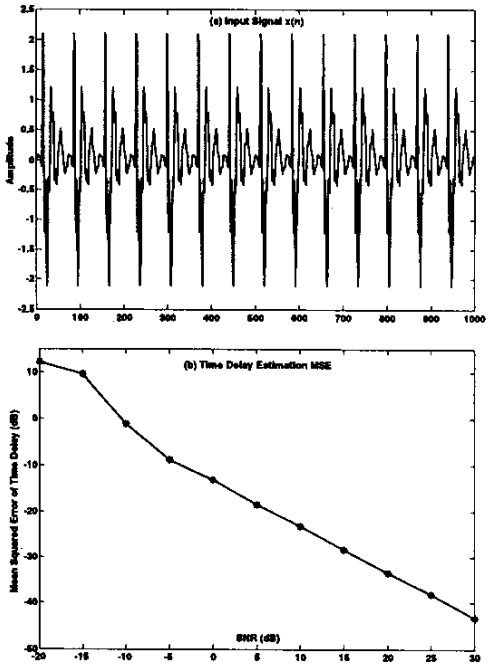

Figure 2: Synthetic Speech "EY"

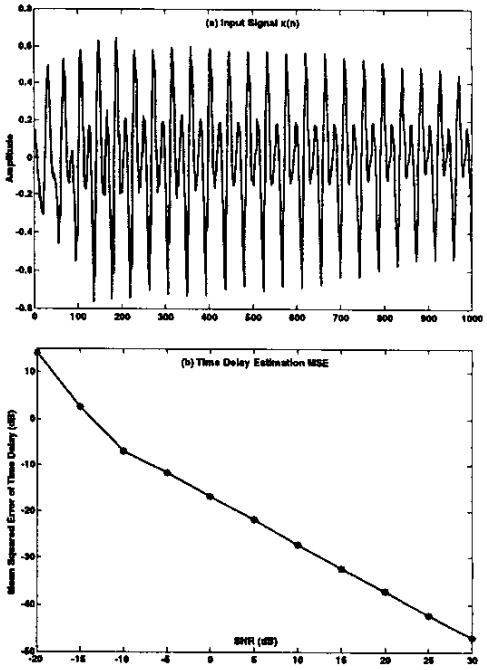

Figure 3: Real Cantonese Speech