SPEECH ANALYSIS BY SUBSPACE METHODS OF SPECTRAL LINE ESTIMATION

Najam Malik and W. Harvey Holmes

School of Electrical Engineering, The University of New South Wales, Sydney, Australia.
N.Malik@ee.unsw.edu.au, H.Holmes@unsw.edu.au.

ABSTRACT

Over frames of short time duration, filtered speech may be described as a finite linear combination of sinusoidal components. In the case of a frame of voiced speech the frequencies are considered to be harmonics of a fundamental frequency. It can be assumed further that the speech samples are observed in additive white noise of zero mean, resulting in a standard signal-plus-noise model. This model has a nonlinear dependence on the frequencies of the sinusoids but is linear in their coefficients. We use subspace line spectral estimation methods of Pisarenko and Prony type to estimate the frequencies and use the results in voiced-unvoiced classification and pitch estimation, followed by analysis of the speech waveform into its sinusoidal components.

1. INTRODUCTION

In [5] McAulay and Quatieri proposed an analysis and synthesis framework based on a model that describes speech, over frames of short time duration, as a finite combination of sinusoidal components. In terms of complex exponentials the model has the form

$$s_n = \sum_{k=-M}^{M} c_k \exp(j\omega_k n) \tag{1}$$

where $c_{-k} = c_k^*$, $\omega_{-k} = -\omega_k$, $\omega_0 = 0$ and $n = 0 \ldots N_s-1$. In the case of a frame of voiced speech the frequencies are considered to be harmonically related to a fundamental frequency ω, $\omega_k = k\omega$. We will assume that the speech samples in (1) are observed in additive white noise of mean zero and variance σ_o^2, resulting in the observed process $x_n = s_n + w_n$, $n = 0 \ldots N_s-1$. The model in (1) has a nonlinear dependence on the frequencies but is linear in the coefficients. In the next two sections we briefly review subspace methods of the Pisarenko and Prony type for estimating the frequencies. We give experimental results in section 4 and in section 5 we discuss the use of these methods in distinguishing between voiced/unvoiced (V/U) frames and in the estimation of pitch. Finally the accuracy of the new pitch estimation algorithm and the suitability of the harmonic sinusoidal model for voiced speech are tested using linear least squares.

2. PISARENKO TYPE SUBSPACE METHODS FOR FREQUENCY ESTIMATION

If we consider the magnitudes and phases of the complex coefficients in (1) to be a system of independent random variables, with the phases uniformly distributed over $(-\pi, \pi)$, then the coefficients c_k are orthogonal. They can also assumed to be uncorrelated with the white noise process w_n. Under these conditions the process $\{x_n\}$ is wide sense stationary and we form a vector from N consecutive random variables of this process, $\mathbf{x} = [x_0\ x_1 \ldots x_N]^T$. By defining the signal vectors $\mathbf{s}_k = [1\ \exp(j\omega_k)\ \exp(j\omega_k 2) \ldots \exp(j\omega_k(N-1))]^T$, for $k = -M \ldots M$, the vectors $\mathbf{c} = [c_M \ldots c_0 \ldots c_M]$ and $\mathbf{w} = [w_0\ w_1 \ldots w_N]^T$, and the matrix $\mathbf{S} = [\mathbf{s}_M \ldots \mathbf{s}_0 \ldots \mathbf{s}_{-M}]$, we can write the observed sequence in matrix-vector form as $\mathbf{x} = \mathbf{Sc} + \mathbf{w}$.

The vectors $\mathbf{s}_M, \ldots, \mathbf{s}_0, \ldots, \mathbf{s}_{-M}$ form a linearly independent set. The $2M+1$-dimensional subspace of $\mathbf{C}^N$ spanned by these vectors is called the signal subspace and its orthogonal complement is called the noise subspace. The correlation matrix of the vector $\mathbf{x}$ has a full set of orthonormal eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N$ and a corresponding set of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$. These eigenvectors and eigenvalues have a special property: The eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{2M+1}$ span the signal subspace and the corresponding eigenvalues are such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{2M+1} > \sigma_o^2$. The remaining eigenvectors

$\mathbf{e}_{2M+2}$, $\mathbf{e}_{2M+3}$, … , $\mathbf{e}_N$ span the noise subspace and their corresponding eigenvalues are all equal, $\lambda_{2M+2} = \lambda_{2M+3} = \ldots = \lambda_N = \sigma_o^2$ [9].

Pisarenko [6] was the first to exploit this structure of the eigenvectors of the correlation matrix in estimating exponentials in noise. A generalisation of the Pisarenko method is the MUSIC procedure (for Multiple Signal Classification) of Schmidt [9]. In the special case of real sinusoids that we are considering here, MUSIC assumes that the number of sinusoids M is known, or can be estimated, and uses a correlation matrix of size $N > 2M+1$. The signal vectors are orthogonal to the noise subspace and hence the inner product of each of these signal vectors with each of the noise subspace eigenvectors is zero. This fact motivates the formation of a general frequency vector $\mathbf{s} = [1 \exp(j\omega) \exp(j\omega 2) \ldots \exp(j\omega(N-1))]^T$ and the definition of the pseudo-spectrum function

$$S_M\left(e^{j\omega}\right) = \frac{1}{\left|\sum_{i=2M+2}^{N} \mathbf{s}^H \mathbf{e}_i\right|} \tag{2}$$

When $\omega = \omega_k$, $\mathbf{s} = \mathbf{s}_k$ and the denominator in (2) is zero. Thus a plot of $S_M(e^{j\omega})$ versus ω would theoretically show infinite peaks at the frequencies ω_k, $-M \le k \le M$.

The denominator in (2) can also be viewed as the evaluation on the unit circle of the polynomial $P_M(z) = \sum_{i=2M+2}^{N} P_i(z)$, where $P_i(z) = e_{i0} + e_{i1}z^{-1} + \ldots + e_{i(N-1)}z^{-(N-1)}$, with e_{ij} the components of $\mathbf{e}_i$. Each of the polynomials $P_i(z)$, and hence the polynomial $P_M(z)$, has a zero on the unit circle at $z = e^{j\omega}$, $\omega = \omega_k$, $-M \le k \le M$. The remaining roots that do not correspond to any signal frequency are called the spurious roots. The summation of the polynomials in $P_M(z)$ has the effect of moving the spurious roots away from the unit circle [9].

A further extension of the above is the minimum-norm method that was put forward by Kumaresan and Tufts [3]. This technique does not directly use all of the vectors in the noise subspace. Instead it selects a single vector $\mathbf{v}$ in the noise subspace which has two properties: the norm of $\mathbf{v}$ is minimum and its first component is 1. Since $\mathbf{v}$ still lies in the noise subspace it is orthogonal to the signal vectors, $\mathbf{s}_k^H \mathbf{v} = 0$ for $-M \le k \le M$, and is thus used to define the pseudo-spectrum

$$S_{MN}\left(e^{j\omega}\right) = \frac{1}{\left|\mathbf{s}^H \mathbf{v}\right|} \tag{3}$$

If the noise subspace eigenvectors are organised into the matrix $\mathbf{E}_n = [\mathbf{e}_{2M+2}\ \mathbf{e}_{2M+3}\ \ldots\ \mathbf{e}_N]$ then $\mathbf{v}$ is given by the expression $\mathbf{v} = \mathbf{E}_n\mathbf{h}/(\mathbf{h}^H\mathbf{h})$, where $\mathbf{h}^H$ is the first row of the matrix $\mathbf{E}_n$. The components of $\mathbf{v}$, $v_0 = 1$, v_1, … , $v_{(N-1)}$, can be used to form a polynomial, $P_{MN}(z) = v_0 + v_1 z^{-1} + \ldots + v_{(N-1)}z^{-(N-1)}$, that again has $2M+1$ roots on the unit circle at frequencies $\omega = \omega_k$, $-M \le k \le M$. The two restrictions on $\mathbf{v}$ force the spurious roots of $P_{MN}(z)$ to be strictly inside the unit circle [3].

3. PRONY TYPE SUBSPACE METHODS FOR FREQUENCY ESTIMATION

Methods of the Prony type use linear prediction in estimation of frequencies. Prony's original method, in the present case of real sinusoids, assumes that noise free observations x_n are available for $n = 0, 1, 2, \ldots, 2(2M+1)-1 = 4M+1$. The signal frequencies can be located by finding the roots of the polynomial $Q(z) = 1 + q_1 z^{-1} + \ldots + q_{(2M+1)}z^{-(2M+1)}$, whose coefficients can be obtained by solving a system of linear equations obtained from the observations. For more details see Scharf [8].

Kumaresan [4] showed that the condition on the number of observations in Prony's method can be relaxed. Define $\mathbf{q}_K = [q_{K0}\ q_{K1}\ \ldots\ q_{KN}]^T$, $Q(z) = q_{K0} + q_{K1}z^{-1} + \ldots + q_{KN}z^{-N}$, and $\mathbf{X}$ to be the coefficient matrix for the homogeneous form of the covariance equations of linear prediction. If the vector $\mathbf{q}_K$ satisfies $\mathbf{X}\mathbf{q}_K = \mathbf{0}$, with $2M+1 \le N \le N_s-(2M+1)$, then the polynomial $Q_K(z)$ has $2M+1$ of its zeros at $e^{j\omega}$, $\omega = \omega_k$, $-M \le k \le M$. In addition the $N-(2M+1)$ spurious roots of the polynomial $Q_K(z)$ are strictly inside the unit circle if its vector of coefficients is chosen such that $q_{K0} = 1$ and its norm, $|\mathbf{q}_K|$, is minimum.

When the measurements are noisy, we can partition $\mathbf{X} = [\mathbf{b}\ \mathbf{A}]$ and $\mathbf{q}_K = [1\ \mathbf{q}'^{\mathrm{T}}_K]^{\mathrm{T}}$ and write the inconsistent system $\mathbf{X}\mathbf{q}_K \approx \mathbf{0}$ as $\mathbf{A}\mathbf{q}'_K \approx -\mathbf{b}$. The matrices $\mathbf{X}$ and $\mathbf{A}$ are generally full rank in the presence of noise. If the number of rows of $\mathbf{A}$, N_s–N, is at least as great as the number of columns, N ($N_s \geq 2N$), then the inconsistent system of equations, $\mathbf{A}\mathbf{q}'_K \approx -\mathbf{b}$, is over determined and it is possible to obtain a least squares minimum norm solution for $\mathbf{q}'_K$ using the pseudo-inverse of $\mathbf{A}$. Such a procedure, however, gives poor frequency estimates. With no noise the rank of the matrices $\mathbf{X}$ and $\mathbf{A}$ is $2M$+1 [4]. In an effort to approach the noise free case, Tufts and Kumaresan [10] proposed the use of a rank $2M$+1 approximation to $\mathbf{A}$ and its pseudo-inverse. Because of this low rank approximation their technique has come to be known as the principle components method. The solution for the coefficient vector uses the singular value decomposition of the matrix $\mathbf{A} = \mathbf{U}_\mathbf{A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}$, and is given by

$$\mathbf{q}'_{KT} = -\left(\mathbf{V}_\mathbf{A}\tilde{\Sigma}^{-1}_\mathbf{A}\mathbf{U}^{\mathrm{H}}_\mathbf{A}\right)\mathbf{b} \tag{4}$$

The matrix $\tilde{\Sigma}^{-1}_\mathbf{A}$ has the inverse of the largest $2M$+1 singular values on its diagonal and zeros elsewhere. The matrix in brackets on the right hand side of equation (4) is the rank $2M$+1 pseudo-inverse of $\mathbf{A}$.

It is possible to simultaneously take into account noise in both $\mathbf{A}$ and $\mathbf{b}$ by employing the method of total least squares of Golub and Van Loan [2], first used for the problem of frequency estimation by linear prediction by Rahman and Yu [7].The total least squares solution $\mathbf{q}'_{TLS}$ is obtained via the singular value decomposition of the data matrix $\mathbf{X} = \mathbf{U}_\mathbf{X}\Sigma_\mathbf{X}\mathbf{V}_\mathbf{X}$, where it is assumed that the number of rows of $\mathbf{X}$, N_s–N, is at least as great as the number of columns, N+1, $N_s \geq 2N$+1. Like the method of principle components, we assume that the singular values σ_j, $j = 2M$+2, …, N+1, arise from noise and are equal. The portion, $\mathbf{V}_n$, of the matrix $\mathbf{V}_\mathbf{X}$, corresponding to these singular values is partitioned as $\mathbf{V}_n^{\mathrm{T}} = [\mathbf{g}^*\ \mathbf{V}'^{\mathrm{T}}_n]$. The minimum norm total least squares solution is then given by

$$\mathbf{q}'_{TLS} = \frac{1}{\mathbf{g}^{\mathrm{H}}\mathbf{g}}\mathbf{V}'_n\mathbf{g}\,. \tag{5}$$

This form of the solution and its equivalence to that obtained by the minimum-norm method was indicated by Dowling and DeGroat [1]. The coefficient vectors $\mathbf{q}_{KT} = [1\ \mathbf{q}'^{\mathrm{T}}_K]^{\mathrm{T}}$ and $\mathbf{q}_{TLS} = [1\ \mathbf{q}'^{\mathrm{T}}_{TLS}]^{\mathrm{T}}$ can be used, in the corresponding polynomials $Q_{KT}(z)$ and $Q_{TLS}(z)$, to define pseudo-spectra analogous to those of section 3.

4. EXPERIMENTAL RESULTS

Figures 1 and 2 show, respectively, a frame of voiced speech and the corresponding pseudo-spectrum obtained by the minimum-norm/total least squares method. Figures 3 and 4 show the same for a frame of unvoiced speech.

The use of the data matrix $\mathbf{X}$ of (4) in total least squares amounts to forming an unnormalised estimate of the correlation matrix by the covariance method. The noise subspace eigenvectors are obtained by singular value decomposition of $\mathbf{X}$. The resulting matrix $\mathbf{V}$ needs to be partitioned by estimating the value of M. In the case of figures 2 and 4 this was done by using forms of the Akaike information criterion and the minimum description length, developed for the case of sinusoids in noise by Wax and Kailath [11]. A simple thresholding of the eigenvalues (singular values squared) can also be employed. Alternatively the dimension of the noise subspace can be fixed (at an under estimate) to accommodate the highest value of M that is likely to be encountered. Doing so would result in a further under estimate of the dimension of the noise subspace when the true value of M is smaller than the maximum. However, the results above would still apply since only fewer noise subspace eigenvectors would be used. The partitioning of the matrix of eigenvectors, regardless of the technique used, implicitly equates the singular values corresponding to the noise subspace eigenvectors. In practice, to accommodate the conditions on N_s, N, and M given in sections 3 and 4, and to keep the computation at a reasonable level, the methods can be employed on subbands.

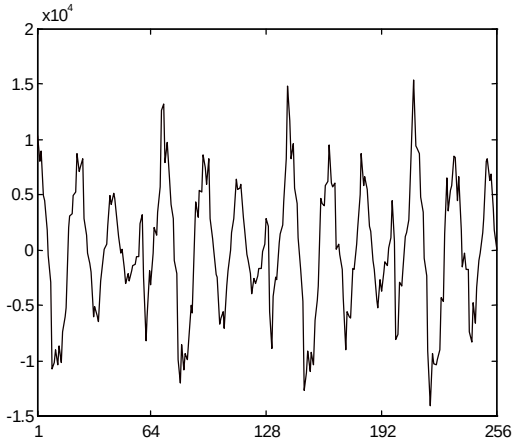

Figure 1: Voiced speech frame.

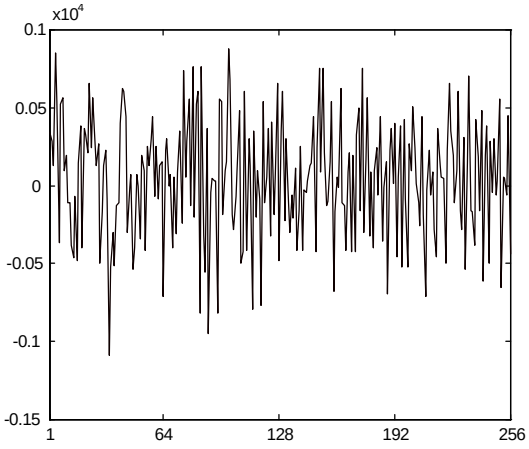

Figure 3: Unvoiced speech frame.

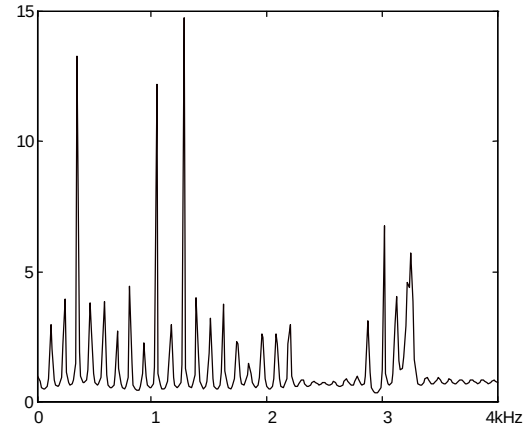

Figure 2: Pseudo-spectrum for frame of figure 1.

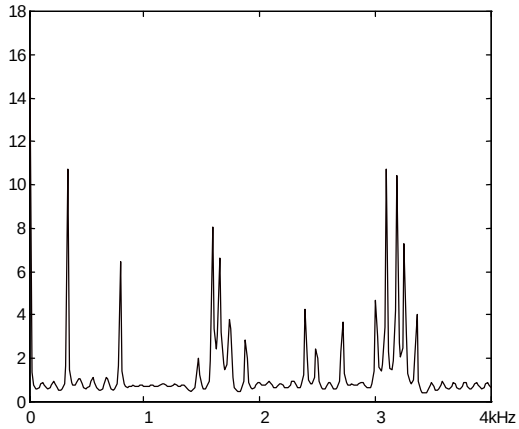

Figure 4: Pseudo-spectrum for frame of figure 2.

5. V/U CLASSIFICATION AND PITCH ESTIMATION

We need to ignore the small ripples in figures 2 and 4 that are due to spurious roots. Examination of the remaining prominent peaks then gives estimates of the frequencies present in the frame of speech under analysis. In voiced speech the frequencies can be modelled as multiples of a fundamental frequency, whereas in unvoiced speech no such relationship can be expected to hold. These assumptions are validated by an examination of figures 2 and 4. The pseudo-spectrum for the frame of voiced speech shows the peak spacing to be quiet uniform. On the other hand the pseudo-spectrum for the unvoiced frame shows the peaks appearing quiet irregularly. This difference can be exploited in V/U classification. A simple decision scheme can be based on the variance of the inter-peak distance together with a threshold. When the variance is small (0.2500 for figure 2) the frame can be considered as voiced and when it is large (1.0612 for figure 4) the frame is declared to be unvoiced.

For pitch frequency estimation the very first peak location can be used. Improved estimates can be obtained by using unweighted and weighted univariate least squares. In these we consider the ith prominent peak location, p_i, to be the ith harmonic of the fundamental/pitch frequency κ, $p_i \approx \kappa i$, and minimise the sum of squared errors, or the sum of weighted square errors. An appropriate set of weights, δ_i for the ith peak, can be obtained from the magnitude of the discrete Fourier transform of the frame of speech being analysed. The assumption that the ith peak is the ith harmonic may not be valid, due to the presence, just inside the unit circle, of spurious roots of the polynomial used in computing the pseudo-spectrum. To estimate pitch frequency, while taking into account the

possibility of missing peaks and nonideal harmonics, we used the following iterative weighted/unweighted least squares algorithm:

$$m_i^0 = i, \quad i = 1,\ldots M_p$$

$$\kappa^j = \frac{\sum_{i=1}^{M_p} \delta_i p_i m_i^{j-1}}{\sum_{i=1}^{M_p} \delta_i \left(m_i^{j-1}\right)^2}, \quad j = 1,2,\ldots$$

$$m_i^j = \text{round}\left(\frac{p_i}{\kappa^j}\right), \quad i = 1,\ldots M_p, \quad j = 1,2,\ldots.$$

where M_p is the total number of peaks and j represents iterations through the algorithm. The final estimate of κ is the value to which the sequence κ^j converges. This convergence takes only a small number of iterations in cases that were tested. The weights δ_i can be set equal to 1 for the unweighted case. It frequently happens that the value M'_p, to which m^j_{Mp} converges, is strictly greater than M_p, $M'_p > M_p$, implying that certain peaks were missing from the pseudo-spectrum. For the case of the voiced frame of figure 1, this algorithm gave the value 114.55Hz. This is very close to estimates taken directly from examination of the frame.

The accuracy of this estimate, and that of the harmonic sinusoidal model for the voiced frame of figure 1, can be evaluated using least squares. Since the matrix **S** of section 1 is now known, standard least squares was used to solve for **c**. The reconstructed frame is shown in figure 5. Its close resemblance to the frame of figure 1 shows that the estimate and the model are very accurate.

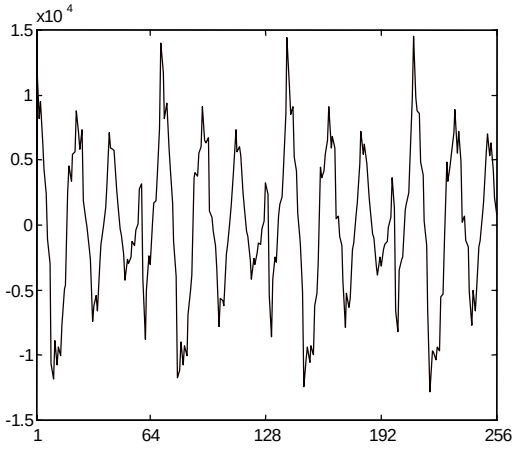

Figure 5: Reconstructed frame corresponding to figure 1.

6. CONCLUSIONS

Subspace methods of frequency estimation are very high resolution spectral estimation techniques that can be applied to speech in the context of the sinusoidal model. Since the methods admit additive noise, we showed that they provide a robust element for V/U classification. In addition, the methods lead to a reliable pitch estimation algorithm.

7. REFERENCES

1. Dowling, E. M., and Degroat, R. D., "The equivalence of the total least squares and minimum norm methods," *IEEE Trans. SP*, vol. 39, pp. 1891-1892, 1991.

2. Golub, G. H., and Van Loan, C. F., “An analysis of the total least squares problem,” *SIAM J. Numer. Anal.*, vol. 17, pp. 883-893, 1980.

3. Kumaresan, R., and Tufts, D. W., “Estimating the angles of arrival of multiple plane waves,” *IEEE Trans. AES*, vol. 19, pp. 134-139, 1983.

4. Kumaresan,R., “On the zeros of the linear prediction error filter for deterministic signals,” *IEEE Trans. ASSP,* vol. 32, pp. 217-220, 1983.

5. McAulay, R. J., and Quatieri, T. F. “Speech analysis/synthesis based on a sinusoidal model,” *IEEE Trans. ASSP*, vol. 34, pp. 744-754, 1986.

6. Pisarenko, V. F., “The retrieval of harmonics from a covariance function*,” Geophys. J. Roy. Astron. Soc.*, vol. 33, pp. 347-366, 1973.

7. Rahman, M. D. A., and Yu, K.-B., “Total least squares approach for frequency estimation using linear prediction,” *IEEE Trans. ASSP*, vol. 35, pp. 1440-1454, 1987.

8. Scharf, L. L., *Statistical Signal Processing: Detection, Estimation and Time-Series Analysis*, Addison-Wesley, Reading, MA, 1991.

9. Schmidt, R., “Multiple emitter location and signal parameter estimation,” *IEEE Trans. AP*, vol. 34, pp. 276-290, 1986.

10. Tufts, D. W., and Kumaresan, R., “Frequency estimation of multiple sinusoids: Making linear prediction perform like maximum likelihood,” *Proc. IEEE*, vol. 70, pp. 975-989, 1982.

11. Wax, M., and Kailath, T., “Detection of signals by information theoretic criteria,” *IEEE Trans. ASSP*, vol. 33, pp. 387-392, 1985.