Robust Cepstral Based Pitch Determination

M. Scott Andrews

Visual Information Technologies
3460 Lotus Dr.
Plano, Texas 75075
(214) 985-2267

Ronald D. DeGroat

Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
Richardson, Texas 75083-0688
(214) 690-2894 degroat@utdallas.edu

Joseph Picone

Speech and Image Understanding Laboratory
Texas Instruments Inc.
P.O. Box 655474 MS 238
Dallas, Texas 75265
(214) 995-6627

Abstract

The FFT based cepstral method of human speech pitch (or fundamental frequency) determination is known to be accurate and reliable in studio quality environments, however, it leaves much to be desired at lower signal to noise ratios. Cepstral pitch determination techniques, which are a special case of the more general theory of homomorphic signal processing, rely on the log operation to deconvolve the pitch sequence from the vocal tract response sequence. Classical cepstral processing models do not account for noise added to the signal. In this paper, we develop a noisy cepstral signal model for speech processing and we propose two Singular Value Decomposition (SVD) based approaches which greatly enhance cepstral based pitch estimation performance in noisy environments.

Speech Production and Cepstral Pitch Determination

Voiced speech production can be modeled reasonably well as a pseudo pulse train (pitch sequence) convolved with a linear system (vocal tract impulse response). Speech is considered wide sense stationary over short time segments (20 - 40 msec) [1] which makes analysis possible over short time windows (or frames). We assume that the z-domain description of the speech signal is modeled by [2], [3]

$$S(z) = H(z)\,P(z) \tag{1}$$

where $H(z)$ is the $z-$transform of the vocal tract response sequence and $P(z)$ is the $z-$transform of the pitch sequence. Analytical expressions for $H(z)$ and $P(z)$ may be found in [2] or [3]. We may use homomorphic filtering techniques to separate the multiplicative relationship in (1) using the complex log operation thereby causing the pitch cepstrum and the vocal tract response cepstrum to occupy approximately disjoint quefrency spaces [2], [4]. Practical implementations of cepstral pitch determination may be obtained from [4] in which it is shown that the Inverse FFT of the log of the magnitude of the FFT provides us with the real version of the quefrency. The connections between the complex cepstrum and the real cepstrum (usually denoted by just cepstrum) are shown in [2] and [3].

The Noise Problem

It is easy to see that homomorphic filtering (cepstral) techniques will not offer good performance in noise. Returning to (1) and taking the complex log operation, we find that

$$\log[S(z)] = \log[H(z)\,P(z)] = \log[H(z)] + \log[P(z)]. \tag{2}$$

The separation of $S(z)$ into its constituent parts works out very neatly assuming that no noise is added to the system. On the other hand, if noise is added to the system, we obtain

$$\log[S(z) + N(z)] = \log[H(z)\,P(z) + N(z)]. \tag{3}$$

A Cepstral Model for Speech Signals in Noise

Manipulating (3) yields a noisy cepstral signal model

$$\log[H(z)\,P(z) + N(z)] = \log[H(z)\,P(z)] + \log\left[1 + \frac{N(z)}{H(z)\,P(z)}\right] \tag{4}$$

which clearly exposes the desired signal component in the first term of the right-hand side. We shall find great utility in going to vector and matrix notation at this point following a discretization of equation (4).

The appropriate discrete Fourier transform (DFT) equivalent of (4) is

$$\log[H(k)\,P(k) + N(k)] = \log[H(k)\,P(k)] + \log\left[1 + \frac{N(k)}{H(k)\,P(k)}\right] \tag{5}$$

where $k = 0, \ldots, M-1$ is the discrete normalized frequency variable. We shall also stay consistent with the notation found in [2] and [3] for representing the log of a general function, $X(k)$, as $\hat{X}(k)$. Thus, we represent (5) in vector form as

$$\hat{\mathbf{x}} = \hat{\mathbf{s}} + \log\left[\mathbf{1} + \mathbf{D}^{-1}\mathbf{n}\right] \tag{6}$$

23ACSSC-12/89/0744 $1.00

where

$$\widehat{\mathbf{s}} = \begin{bmatrix} \log[H(0)P(0)] \\ \log[H(1)P(1)] \\ \vdots \\ \log[H(M)P(M)] \end{bmatrix}_{M\times 1} \tag{7}$$

$$\mathbf{n} = \begin{bmatrix} N[0] \\ N[1] \\ \vdots \\ N(M) \end{bmatrix}_{M\times 1} \tag{8}$$

$$\mathbf{D} = diag\left[H(0)P(0), H(1)P(1), \cdots, H(M)P(M)\right]_{M\times M}. \tag{9}$$

We shall refer to various aspects of this signal model as we bring out different results from our preliminary investigations.

SVD Based Cepstral Estimates - The MUSIC Approach

Schmidt [5] in 1979 formally proposed an SVD based algorithm from which to estimate the power spectral density (PSD) of time series observations emitted by sets of independent sources. Of course our application and use of the algorithm is in a much different framework, nevertheless it still applies.

If our data in (6) is structured into a Toeplitz data matrix, then a wide variety of techniques can be used to detect the simple pitch harmonic in the quefrency domain. Pioneering work in using the SVD (eigenvector method) has been done by Tufts and Kumaresen [11] in estimating signal parameters of noise speech-like signals. The MUSIC method is one such method which employs the SVD for estimating the Power Spectral Density (PSD) of a signal. It may be written compactly as [6]:

$$\widehat{\Psi}_{MUS}(\omega) = \frac{1}{\mathbf{W^H}\sum_{i=p+1}^{N}\mathbf{U_i U_i^H W}} \tag{10}$$

where

$$\mathbf{W} = \frac{1}{\sqrt{N}}\left[1\ e^{j\omega}\ e^{j2\omega}\ \cdots\ e^{j(N-1)\omega}\right]^t \tag{11}$$

and $\mathbf{U_i}$ is the ith singular vector of the data matrix $\mathbf{X} = \mathbf{U\Sigma V^t}$. This technique essentially uses the estimated noise subspace, given by $\{\mathbf{U_{p+1}}, \mathbf{U_{p+2}}, \cdots, \mathbf{U_N}\}$ to obtain a high resolution PSD.

The MUSIC method works well with narrow-band signals in additive white Gaussian noise. In our application, the noise is approximately white within a given data vector, but the noise power (and the signal to noise ratio (SNR)) varies from one vector to the next. We therefore propose a technique which will tend to equalize the noise power and at the same time de-emphasize vectors with lower SNRs.

Noise Equalization via Vector Normalization

We start with an original data matrix, $\mathbf{X}$ (dimension $M \times N$), structured in a Toeplitz form. We then represent our data matrix as a set of N rows of $M \times 1$ column vectors by the notation

$$\mathbf{X} = [\mathbf{x_1}|\mathbf{x_2}|\cdots|\mathbf{x_N}]. \tag{12}$$

If we form a diagonal matrix, $\mathbf{D}$, from the column vector norms of $\mathbf{X}$ in (12), i.e.

$$\mathbf{D} = diag\left(\|\mathbf{x_1}\|, \|\mathbf{x_2}\|, \cdots, \|\mathbf{x_N}\|\right) \tag{13}$$

then we may form a normalized output matrix, $\mathbf{Y}$, from $\mathbf{X}$ and $\mathbf{D}$ as

$$\mathbf{Y} = \mathbf{XD^{-1}}. \tag{14}$$

The $\mathbf{D}$ matrix acts as a primitive whitening filter on the original data matrix, $\mathbf{X}$. If $\widehat{\mathbf{x}}_i = \widehat{\mathbf{s}}_i + \widehat{\eta}_i$ where $\widehat{\eta}_i$ is the "noise" component corresponding to $\log\left(\mathbf{1} + \mathbf{D^{-1}n}\right)$, then $\|\widehat{\mathbf{x}}_\mathbf{i}\| \doteq \|\widehat{\eta}_\mathbf{i}\|$ at low $SNR(< 0dB)$. Normalizing the data vectors as shown in (14) will tend to equalize the noise power in each vector. This technique will also tend to de-emphasize the lowest SNR vectors. If equalization is applied to the data before MUSIC, we will call it equalized MUSIC or E-MUSIC.

Before leaving this section it is worth mentioning that other variants of the SVD based PSD estimator exist. One such variant is the eigenvector method of Tufts and Kumaresan [11], [13], and [14] which is similar to the MUSIC method except that the singular vectors used to obtain the spectral estimates are weighted by their respective singular values, σ_i.

Simulation Results Using Real Speech Data

In this section we show our results from simulations using the Texas Instruments (TI) pitch database mentioned in [8]. The overall characteristics, in terms of signal-to-noise ratio, of the database are shown in Figure 1 [8]. Data were obtained over telephone lines, thus introducing a variety of channel conditions into the original speech. Some of these channel conditions were quite difficult to model. Reference files were generated to optimize synthetic speech quality [8]. These reference files were used to evaluate the performance of the pitch determination algorithms on the database.

Figures 2, 3, 4, and 5 show liftered (filtered in cepstral domain, after [9]) cepstrums in the range of the desired pitch peak. Voiced and unvoiced frames were excised from the TI pitch database. Standard FFT cepstral processing methods, as well as our new FFT-MUSIC cepstral methods, were used to generate the cepstrums of the representative voiced and unvoiced frames, respectively. Clearly, the FFT-MUSIC cepstral method finds a distinct peak in the correct location for the pitch, whereas the FFT-FFT cepstral generator fails.

Figures 6 and 7 show dependence on window length of gross pitch error given that we know a priori that the signal was voiced. This dependence was predicted by Verhelst and Steenhaut, in [10], specifically for the cepstrum case. For the more general case, this dependence was predicted by Picone et. al.,

in [11]. In Figure 6 we show, for a correlation matrix column vector length of 180 (from a 240 sample space corresponding to 30 msec of speech), and taking 14 singular vectors as "signal" subspace vectors, performance for the regular MUSIC cepstral method and our E-MUSIC cepstral method. The E-MUSIC method has statistically better performance associated with it when other parameters are not adjusted optimally. It suffices to say that since the noise spread of the data base was never less than 5 dB SNR, we were unable to verify our E-MUSIC cepstral method. Our E-MUSIC method, through preliminary investigations, does not show significant improvements over MUSIC for signals greater than 0 dB SNR. However, when the SNR is at or slightly below 0 dB, we have seen significant improvements with E-MUSIC over MUSIC on a number of test cases. We show, in Figure 7, an optimal window length found at the trough for the above parameters, using the FFT-MUSIC cepstral processing method. This optimal was found to be 11 msec and was used in a number of simulations.

For a correlation matrix column vector length of 180 and a fixed long window length of 64, we show, in Figure 8, the error dependence on the number of "signal" subspace vectors. The MUSIC and E-MUSIC methods are compared. Again, statistical outperformance of the E-MUSIC in this case is due to unoptimally adjusted parameters (in this case window length). These curves approach each other as other parameters become optimally adjusted. Also, we should note that it is no accident that the optimal number of signal subspace vectors is at 14. It turns out that 12 is the optimal LPC order to use for processing this database.

In Figure 9, we show consecutive cepstra from our FFT-MUSIC method for one of the more difficult speech files in the database. We compare this to the FFT-FFT generated cepstra in Figure 10. From the figures we see that a more clearly visible pitch track is observed from the MUSIC generated cepstra. The gross pitch error, given a priori that a frame is voiced, was reduced to 3.11 % for the entire database of 3 minutes of speech from 38 speakers. The standard FFT-FFT cepstral method produced a gross pitch error of 26.14 %. This performance improvement demonstrates a significant enhancement to deconvolution problems, in general, when we are processing noisy signals.

Finally, Figure 11 shows our "optimality surface" for our two parameters of correlation matrix vector length and number of signal subspace vectors. The window length for the data was fixed at our discovered optimal value of 11 msec.

Conclusions

We have looked at a new technique for pitch determination in the quefrency domain using a modified SVD based approach. We have shown that significant enhancements to cepstral pitch determination are possible with the FFT-MUSIC cepstral processing method. Furthermore, we have demonstrated statistical improvement, for unoptimally adjusted parameters, of the E-MUSIC method, introduced in this paper, over the MUSIC method.

Since traditional cepstral pitch determination techniques involve deconvolution, we have also demonstrated performance enhancements to deconvolution problems, in general, using our FFT-MUSIC cepstral generation methods. We expect this improvement to have impacts in other disciplines as well.

We expect we can improve on our results by integrating the FFT-MUSIC cepstral processing method with other algorithms which depend on accurate cepstral coefficient generation. Also, we can improve our results by integrating our method with more intelligent algorithms such as dynamic programming optimization algorithms and Markov models.

Since many of the current state-of-the-art pitch tracking techniques depend on LPC, we believe that by lowering the SNR characteristics of the database, we will outperform such techniques. Such outperformance is simply due to the fact that the threshold breakdown for the MUSIC technique (based on the SVD) is lower than LPC techniques. This will especially be true if we use our E-MUSIC method in place of MUSIC for getting at the cepstrum.

Finally, we note that combining appropriate PSD methods produces a whole family of possible algorithms for deconvolution with different levels of performance and computation. We will investigate other combinations in the future.

References

[1] T. Parsons, Voice and Speech Processing, Mcgraw-Hill Book Company, New York, 1986.

[2] A.V. Oppenheim and R.W. Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, 1975.

[3] A.V. Oppenheim and R.W. Schafer, "Homomorphic Analysis of Speech," IEEE Trans. Audio Electroacoust., Vol. AU-16, No. 2, June 1968, pp. 221-226.

[4] A.M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Amer., Vol. 41, Feb. 1967, pp. 293-309.

[5] R.O. Schmidt, "A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation," Ph.D. dissertation, Stanford University, 1981.

[6] S.M. Kay and C. Demeure, "The High-Resolution Spectrum Estimator - A Subjective Entity," Proc. IEEE, Vol. 72, no.12, pp. 1815-1816, December 1984.

[7] J.D. Markel and A.H. Gray, Jr., Linear Prediction of Speech, Springer-Verlag, New York, 1976.

[8] J. Picone, G. Doddington, B.G. Secrest, "Robust Pitch Detection in a Noisy Telephone Environment," in Proc. IEEE ICASSP, 1987, pp. 1442-1445.

[9] B.P. Bogert, M.J.R. Healy, J.W. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-autocovariance, Cross-Cepstrum, and Saphe Cracking," Proc. Symp. Time Series Analysis, M. Rosenblatt, Ed., New York, John Wiley & Sons, Inc., New York, 1963, pp. 209-243.

[10] W. Verhelst and O. Steenhaut, "A New Model for the Short-Time Complex Cepstrum of Voiced Speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, pp. 43-51, February, 1986.

[11] J. Picone, D.P. Prezas, W.T. Hartwell, and J.L. Locicero, "Spectrum Estimation Using an Analytic Signal Representation," Signal Processing, North-Holland, Vol. 15, no. 2, pp. 169-182, September 1988.

Figure 1: Noise Characteristics of Pitch Database

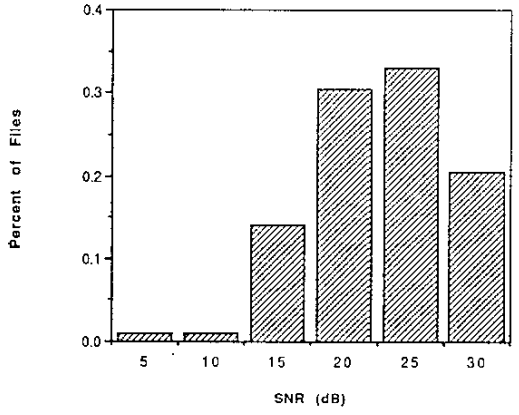

Figure 2: FFT-FFT Cepstrum of Representative Unvoiced Frame

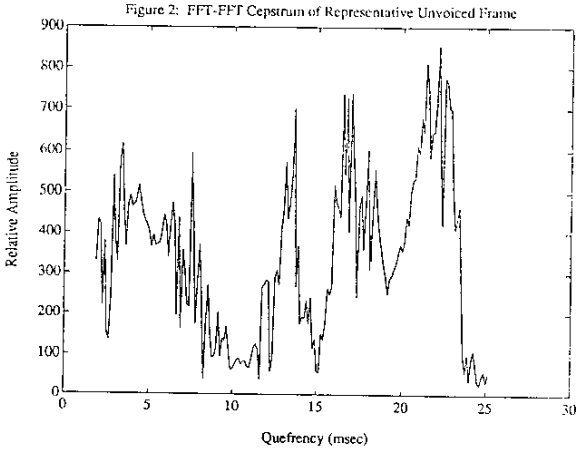

Figure 3: FFT-MUSIC Cepstrum of Representative Unvoiced Frame

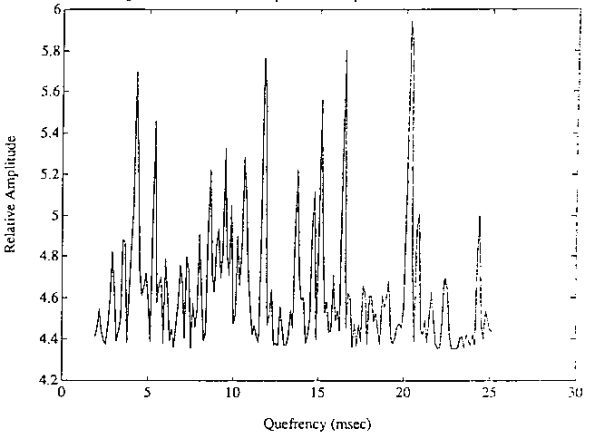

Figure 4: FFT-FFT Cepstrum of Representative Voiced Frame

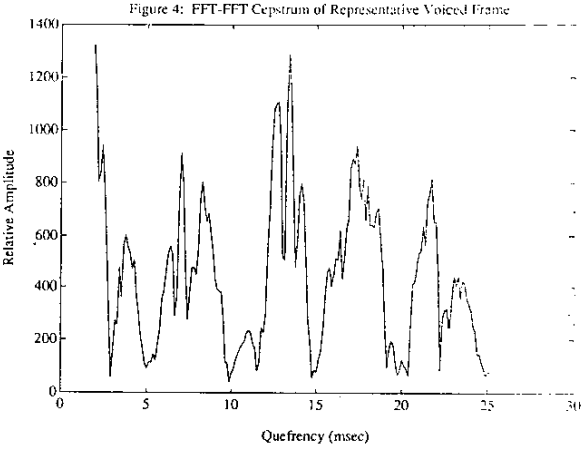

Figure 5: FFT-MUSIC Cepstrum of Representative Voiced Frame

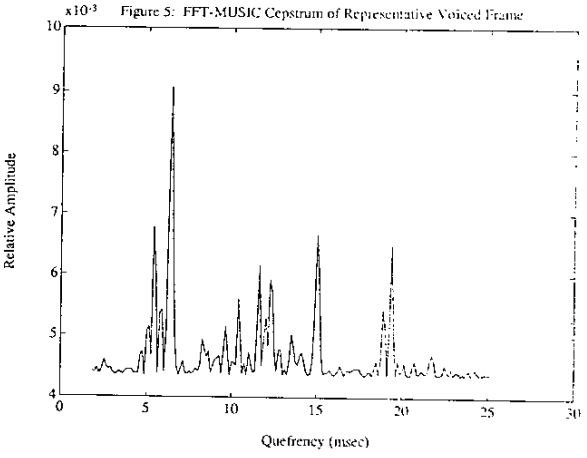

Figure 6: Error Dependence on Window Length (dashed = E-MUSIC)

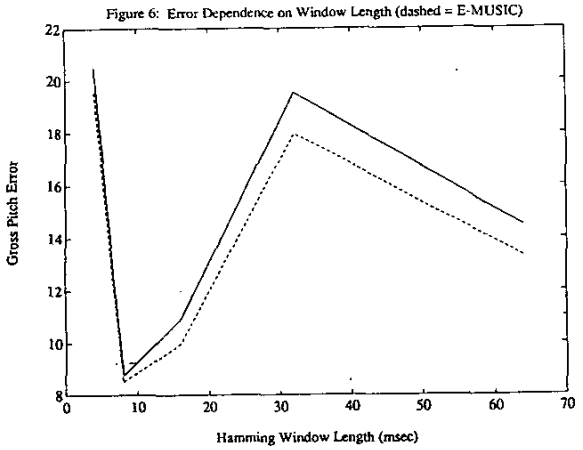

Figure 7: Error Dependence on Window Length (FFT-MUSIC Cepstra)

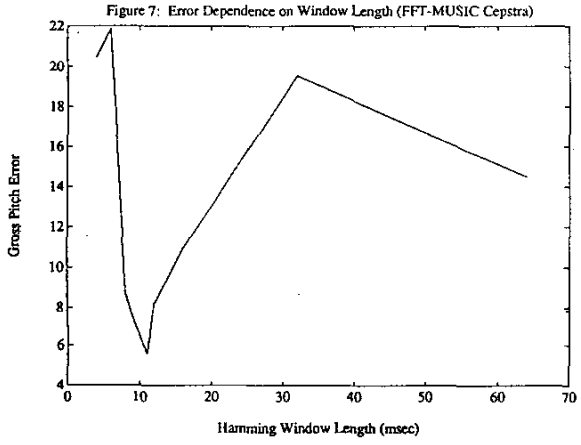

Figure 8: Error Dependence on Singular Values (dashed = E-MUSIC)

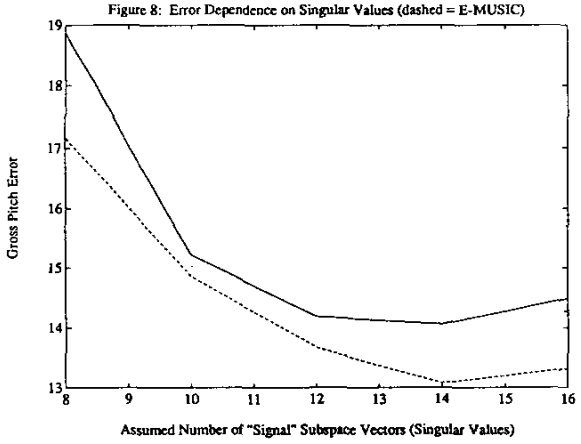

Figure 9: FFT-MUSIC Generated Cepstra for 86 Consecutive Frames

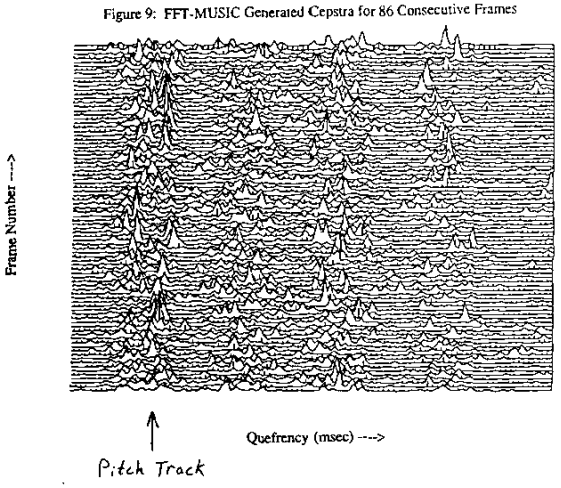

Figure 10: FFT-FFT Generated Cepstra for 86 Consecutive Frames

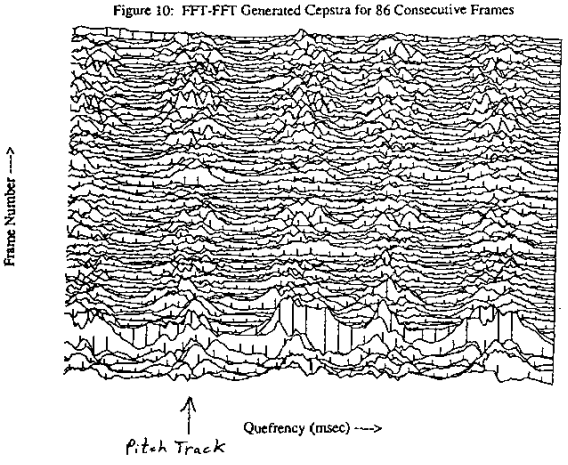

Figure 11: Optimality Surface for Fixed 11 msec Window