A High Performance MUSIC Based Pitch Detector

M.S. Andrews
Comar Inc.
1800 N. Glenville, #100
Richardson, TX 75081
(214) 238-7691
andrews@utdallas.edu

R.D. DeGroat
Erik Jonsson School of Engineering
The University of Texas at Dallas
Richardson, TX 75083-0688
(214) 690-2894
degroat@utdallas.edu

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

Robust fundamental frequency detection and estimation of a speech signal has remained an elusive goal in speech processing; yet, there will always be a need for reliable detection and estimation of this important speech feature in adverse environments. Heuristic approaches to the problem have been optimized for specific environments [1, 2]; but such approaches cannot, in general, be applied to other environments where variations in channel conditions, background noise, and transduction processes abound.

In this paper, we extend some of the work previously reported [3, 4] on a greatly enhanced cepstral based pitch detector and estimator which employs the MUSIC [5] algorithm. We demonstrate the ability of subspace processing [6, 7] to not only perform excellent pitch estimation, as judged by objective performance evaluations, but also to perform good classification of voiced/unvoiced speech signals. Furthermore, our techniques have shown marked strides of improvement in estimation at very low signal-to-noise ratios [4].

1. Introduction and Background

The FFT based cepstral method, since its inception [8], has been considered to be an accurate and reliable method for determining fundamental frequency of a speech signal, provided that the signal was produced in a clean environment (typically referred to as studio quality data). Two drawbacks to this historical technique are that it leaves the detection problem unanswered and it does not handle additive noise.

The classical cepstrum can be succintly described as an FFT-log-FFT operation. In [3], we introduced an FFT-log-MUSIC based cepstral algorithm for estimating pitch in a speech signal. MUSIC [5] is a high resolution spectral estimator. The Texas Instruments Long Distance Telephone Pitch Detection database of [1] was used to evaluate the performance. We showed significant improvement in estimation performance over standard FFT–log–FFT based processing [8] given that we knew a priori that the speech signal was voiced. This a priori knowledge came to us from the hand edited reference pitch tracks of [1] which were edited in such a way so as to optimize synthetic speech quality. We found, that when tested against such objective measures, our method had a 3.11% pitch estimation error compared against a 26.14% estimation error for the classical FFT based method. We have reached as low an estimation error as 2.69% in further studies using the FFT-log-MUSIC technique introduced in [3].

Our first attempt at placing the FFT–log–MUSIC estimator in [3] in an operational environment is reported in [4]. There we show that at low signal-to-noise ratios, our estimation performance is significantly better than all other known methods in the literature. One of the problems that we encountered in moving this improved MUSIC based cepstral estimator to an operational environment was the lack of a voicing decision mechanism, other than the use of the dynamic programming based tracking of [1]. So-called high resolution frequency estimators of the MUSIC class suffer from poor amplitude estimation abilities and hence we arrived at poor classification of voiced/unvoiced speech signals.

In this paper, we show some improvements to our previously reported techniques by using a MUSIC-log-MUSIC operation to attain a *superresolution cepstrum*. Since we rely on the singular value decomposition (SVD) of a data matrix to get our spectral and cepstral estimates, we can use the singular values to make preliminary voicing decisions. It turns out, that from our experimentation, various tests involving the singular values provide varying degrees of discrimination between the voiced and unvoiced speech data. We also mention here that there still remain some difficulties with the SVD tests in transitional speech data - speech which is moving in to and out of voiced regions and in to and out of unvoiced regions. We will detail our results with appropriate explanations following a brief review of cepstral methods for pitch analysis.

2. Classical and Modern Cepstral Methods for Pitch Analysis

Classical cepstral methods rely on the simplified z-domain description of a speech signal, assumed to be a model of the form:

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

where $H(z)$ is the $z-$transform of the vocal tract response sequence and $P(z)$ is the $z-$transform of the glottal excitation (or pitch sequence). More detailed analytical expressions, based on some simplifying assumptions of Eq. 1, may be found in [8, 9]. What is important is that $P(z)$ takes on the form of a periodic, or quasi-periodic, pseudo-pulse train when the speech signal is voiced. When the speech signal is unvoiced, $P(z)$ can be assumed to take on the guise of the z-transform of a white noise sequence.

In cepstral processing, it is desired to separate $P(z)$ from $H(z)$ in Eq. 1 by converting the multiplicative relationship into

0-4ACSSC-12/90/0669 $1.00

an additive one. This can be accomplished with the complex log operator, i.e., $\log[H(z)P(z)] = \log[H(z)] + \log[P(z)]$. This operation, however, presents a problem (from an algebraic standpoint) when there is an additive component to Eq. 1, such that in the $z-$domain, we obtain:

$$\begin{aligned}\log[S(z)+N(z)] &= \log[H(z)P(z)+N(z)] \\ &= \log[H(z)P(z)] + \log\left[1+\frac{N(z)}{H(z)P(z)}\right] \end{aligned} \quad (2)$$

In [3] we first introduced a discretized vector form of Eq. 2, which exposes once again the desired signal component that we seek. This expression is repeated again below for convenience:

$$\hat{\mathbf{x}} = \hat{\mathbf{s}} + \log\left[\mathbf{1} + \mathbf{D}^{-1}\mathbf{n}\right] \quad (3)$$

where

$$\hat{\mathbf{s}} = \begin{bmatrix} \log[H(1)P(1)] \\ \vdots \\ \log[H(M)P(M)] \end{bmatrix}_{M\times 1} \quad (4)$$

$$\mathbf{n} = \begin{bmatrix} N[1] \\ \vdots \\ N(M) \end{bmatrix}_{M\times 1} \quad (5)$$

$$\mathbf{D} = diag[H(1)P(1),\cdots,H(M)P(M)]_{M\times M}. \quad (6)$$

It is the presence of the noise term (Eq. 5) in Eq.3 that produces undesirable behavior in the cepstrum. As explained in [3], at low SNR (< 10 dB), the entire second term in Eq. 3 turns out to be approximately white. It is under this case that the MUSIC algorithm performs its best where cepstral processing is concerned. As indicated in [4], it is also under this case that the MUSIC based cepstral algorithm performs better than all others where the pitch estimation problem is concerned.

3. MUSIC Based Cepstral Processing - The Superresolution Cepstrum

In [3], we described a family of possible algorithms that could be used to obtain high resolution cepstral estimates. In an attempt to improve our estimation performance of the pitch harmonic, we now apply a MUSIC–log–MUSIC approach as another algorithm in this class. We obtained significant reductions in pitch estimation error over standard FFT–log–FFT methods and over our FFT–log-MUSIC method.

The double MUSIC algorithm consists of a MUSIC front end, a log operation and a MUSIC back end, which produces a superresolution cepstrum (MUSIC is used in place of both FFTs). The first step is to form a Toeplitz data matrix from the N data points,

$$\mathbf{X} = \begin{bmatrix} x(P) & x(P-1) & \cdots & x(1) \\ x(P+1) & x(P) & \cdots & x(2) \\ \vdots & & & \vdots \\ x(N-1) & x(N-2) & \cdots & x(N-P) \end{bmatrix}_{(N-P)\times P} \quad (7)$$

for which the singular value decomposition (SVD) is

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^t \quad (8)$$

The size of the data matrix is determined by $P \leq N-P$ which is chosen to optimize performance over the T.I. database. If we stay with compact notation (introduced in [10]) used to describe the MUSIC spectrum, we have the following results:

$$\Psi_{MUS}(\omega) = \frac{1}{\mathbf{W^H}\sum_{i=Q+1}^{N-P}\mathbf{U_i}\mathbf{U_i}^H\mathbf{W}} \quad (9)$$

where

$$\mathbf{W} = \frac{1}{\sqrt{N-P}}\left[1\ e^{j\omega}\ e^{j2\omega}\ \cdots\ e^{j(N-P-1)\omega}\right]^t \quad (10)$$

and U_i is the ith singular vector of the data matrix $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^t$. This technique essentially uses the estimated noise subspace, given by $\{U_{Q+1}, U_{Q+2}, \cdots, U_{N-P}\}$ to obtain a high resolution spectral estimate. Next we form the log of Eq. 9, as

$$\hat{\Psi}_{MUS}(\omega) = \log[\Psi_{MUS}(\omega)]. \quad (11)$$

Now if we appropriately discretize Eq. 11 and use the data to form the log-spectral covariance matrix, we obtain the Toeplitz structure

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{\Psi}(p) & \hat{\Psi}(p-1) & \cdots & \hat{\Psi}(1) \\ \hat{\Psi}(p+1) & \hat{\Psi}(p) & \cdots & \hat{\Psi}(2) \\ \vdots & & & \vdots \\ \hat{\Psi}(n-1) & \hat{\Psi}(n-2) & \cdots & \hat{\Psi}(n-p) \end{bmatrix}_{(n-p)\times p} \quad (12)$$

from which we take the singular value decomposition of Eq. 12 as

$$\hat{\mathbf{Y}} = \hat{\mathbf{U}}\hat{\Sigma}\hat{\mathbf{V}}^t \quad (13)$$

then we can form our new superresolution cepstral estimation function as

$$\mathrm{X}_{MUS}(\tau) = \frac{1}{\tilde{\mathbf{W}}^H\sum_{i=q+1}^{n-p}\hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^H\tilde{\mathbf{W}}} \quad (14)$$

where

$$\tilde{\mathbf{W}} = \frac{1}{\sqrt{n-p}}\left[1\ e^{j\tau}\ e^{j2\tau}\ \cdots\ e^{j(n-p-1)\tau}\right]^t. \quad (15)$$

Most of the time, the simple choice of the maximum in the function expressed in Eq. 14 will yield an excellent estimate of the pitch, given that the frame of speech data over which this function is applied is clearly voiced. The only problem with this is that by the time we have applied the MUSIC algorithm twice over the data matrix, with a log operation in between, the amplitude information is virtually lost. There is little hope of obtaining a normalized version of this function which is representative of the true energy (or partial energy) in the original data. This poses no problem, however, if we consider that the energy information that we desire for such a detection problem as this is contained in the squares of the singular values of the first decomposition.

4. Singular Value Based Pitch Detection

The square of the singular values from the first SVD (eq. 8) provide us with an energy measure of the data that leads to a general discriminant for voicing decisions:

$$\frac{S_{ij}}{S_{kl}} = \frac{\sum_{m=i}^{j} \sigma_m^2}{\sum_{m=k}^{l} \sigma_m^2} \quad > T \tag{16}$$

where T is some chosen threshold which when exceeded by the ratio will cause a choice to be made in favor of a voiced speech frame and i, j are chosen to correspond to the "signal" subspace and k, l are chosen to correspond to the "noise" subspace. We may view the ratio as an attempted classification of independent distributions for voiced and unvoiced speech. This implies that in the case of narrowband signals (or their equivalent) in the presence of broadband noise, the ratio S_{ij}/S_{kl} will tend to be larger than in the case of just background noise.

A variant of the above ratio test which we have experimented with involves setting of the denominator in the ratio of Eq. 16 to unity and $i = j = 1$. This action happens to be the equivalent of taking the two-norm squared of the data covariance matrix, expressed as

$$||X||_2^2 = \sigma_1^2 \tag{17}$$

such that for a voiced speech data frame, we have

$$||X||_2^2 = \sigma_1^2 > T. \tag{18}$$

Our preliminary investigations indicate that eq. 18 can be used to reduce the possibility of choosing an unvoiced frame as voiced by nearly 50% and this is considered good discrimination by many speech reasearchers. There are other suboptimal methods available, all based on Eq. 16 and each has its specific meaning as far as signal model is concerned. In the case of Eq. 18, we are looking at rank one energy. We have also found good success with the choice of a 10 dimensional "signal" subspace test, where $i = 1, j = 10, k = 1, l = P$, and P is the minimum dimension of the data matrix. This test is tantamount to a normalized energy ratio of the energy in the "signal" subspace (with at most $10/2 = 5$ narrowband processes) and the total energy in the frame.

5. Results of Simulations and Evaluations on an Operational Database

Figure 1, in conjunction with Table 1, summarizes appropriate statistics of the Texas Instruments Long Distance Pitch Detection database [1] used to test the MUSIC based cepstral pitch algorithm. First we turn to the estimation problem and observe some typical results. Figures 2, 3 and 4 taken together reveal our progress in the area of cepstral estimation. We can view classical FFT based methods from Figure 2. We note that very poor estimation performance is clear by observation. Peaks which remotely resemble pitch locations are biased and inconsistent (an historical problem with FFT-log-FFT processing). We see from Figure 3 a significant improvement in both peak location and consistency, when we use the FFT-log-MUSIC approach. Our best achievement, as far as estimation is concerned, comes from the MUSIC-log-MUSIC method. We see that there are few choices as far as picking out a pitch peak is concerned. The pitch estimation error for the entire database is on the order of 1% of all detected voiced frames.

We now turn our attention to the detection problem. Our previous performance as reported in [4], was based on the amplitude output by the MUSIC function. The a posteriori distributions of the MUSIC amplitude function for this database is given in Figure 5. We observe that reliance on the MUSIC amplitude function translates to very poor separability in the distributions. Since we have implemented a two pass MUSIC approach for determining the cepstrum, as mentioned previously, we can use the information provided by the singular values of the first pass (indicating signals present or absent in the power spectrum) to assist our voicing decision problem. We already have the use of a dynamic programming based tracking algorithm helping us make decisions based somewhat on the amplitude of the pitch peak and the consistency in frequency with which it occurs. Mainly due to the inconsistency of MUSIC amplitude, we rely more on the consistency of the frequency estimate. Now since our singular values tell us something about voicing, we use this knowledge to input a zero-cepstrum to the dynamic tracking algorithm. This action assures the clearly unvoiced frames will not even enter into the decision making process. In Figures 6-8, we show progressively improved speech data in terms of SNR. We show that at low SNR, separability based on eigenvalues becomes more difficult, as expected. Observation of Figures 7 and 8 show that as we increase SNR, we improve our separability. In these figures, we have shown a surface plot of the 120 singular values, obtained from the SVD of the Toeplitz data matrix (Eq. 8), against consecutive voiced frames and then consecutive unvoiced frames, concatentated on to the same surface for ease of comparison. One of the big problems with high resolution spectral estimation (and hence also with superresolution cepstral estimation) is spurious peaks allowed in the cepstrum through inexact modeling. Utilizing knowledge of whether or not the cepstrum should even have a peak (representing pitch) before computing it is useful for improving the modeling process. Empirically derived distributions of the log of the eigenvalues for the case of the Texas Instruments Long Distance Telephone Pitch Detection Database are shown in Figure 9. We took the log to limit the range in the distributions for visual display purposes. The classical detection problem is apparent by inspection of the distributions. In the case of our algorithms, and in consideration of the overall pitch tracking system, we choose to favor the voicing distribution such that we end up eliminating about 50% of all unvoiced frames from consideration. The other 50% of all unvoiced frames includes all transitional frames (in to and out from voiced segments) and channel dominated frames (in which a colored noise may be present), both of which will likely contain sufficiently high variances to favor a positive voicing decision to be hypothesized.

For this database, an optimal T was discovered at $2.0982x10^8$. We applied the two pass MUSIC with this T and have obtained a cumulative performance of 6.96%. This compares favorable to current state of the art, Texas Instruments integrated correlator [1], which had a cumulative performance of 3.82%.

6. Concluding Remarks

We have employed the use of the superresolution cepstrum in order to achieve the goals of state-of-the-art in pitch detection and estimation. The current positioning of this method with other methods studied in the literature [1] can be viewed from Figure 10. In terms of error from synthetic speech quality optimizations, this Figure shows the relative performance in comparison with other popular pitch detection and estimation algorithms. Although we have not achieved state-of-the-art using the superresolution cepstral techniques described herein, we recognize that it is certainly approachable and probably within grasp of these methods.

Our plans for continued investigation of these techniques applied to the problem of pitch detection and estimation include a rigorous statistical analysis of the algorithm, application of minimum norm solutions to the problem, least squares amplitude estimation, and further study into the optimal choices of i, j, k, l, and T of Eq. 16. These appear to be good topics for further investigations into what we have presented here.

Bibliography

[1] J. Picone, G. Doddington, and B. G. Secrest. Robust pitch detection in a noisy telephone environment.

[2] J. P. Campbell and T. E. Tremain. Voiced/Unvoiced Classification of Speech with Applications to the U.S. Government LPC-10E Algorithm. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 473–475, Tokyo, Japan, April 1986.

[3] M. S. Andrews, R. D. DeGroat, and J. Picone. Robust cepstral based pitch determination. In *Proceedings 23rd Asilomar Conference on Signals, Systems, and Computers*, pages 744–748, Pacific Grove, California, October 1989.

[4] M. S. Andrews, J. Picone, and R. D. DeGroat. Robust pitch determination via svd based cepstral methods. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 254–256, Albuquerque, New Mexico, April 1990.

[5] R. O. Schmidt. *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*. PhD thesis, Stanford University, 1981.

[6] D. W. Tufts and R. Kumaresan. Improved spectral resolution ii. In *Proceedings IEEE ICASSP*, pages 592–597, 1980.

[7] R. Kumaresan and D. W. Tufts. Accurate parameter estimation of noisy speech-like signals. In *Proceedings IEEE ICASSP*, 1982.

[8] A. M. Noll. Cepstrum pitch determination. *Journal Accoust. Soc. America*, 41:293–309, February 1967.

[9] W. Verhelst and O. Steenhaut. A new model for the short-time complext cepstrum of voiced speech. In *IEEE Trans. Acoust., Speech, Signal Processing*, volume ASSP-34, pages 43–51, February 1986.

[10] S. M. Kay and C. Demeure. The high-resolution spectrum estimator - a subjective entity. In *Proceedings IEEE*, volume 72, 1984.

Table 1: Characteristics of Texas Instruments Long Distance Telephone Pitch Detection Database

Total Speech Time:	3 Minutes
Total Utterances:	60
Total Adult Male Utterances:	39
Total Adult Female Utterances:	21
Percent of Pitch Frequencies Under 150 Hz:	62 %
Percent of Pitch Frequencies Under 250 Hz:	32 %
Percent of Pitch Frequencies Under 350 Hz:	5 %
Percent of Pitch Frequencies Over 350 Hz:	1 %

Figure 1: SNR Distribution of TI Long Distance Pitch Database

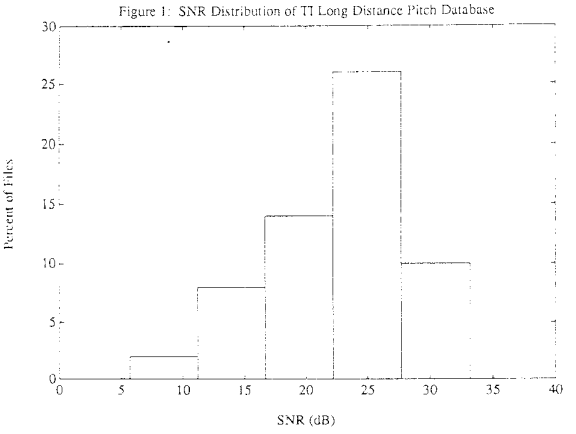


Figure 2: FFT-FFT Cepstra of Speech File with SNR = 9.5 dB

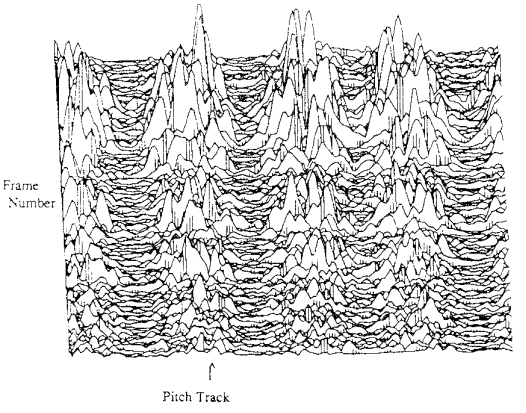


Figure 3: FFT-MUSIC Cepstra of Speech File with SNR = 9.5 dB

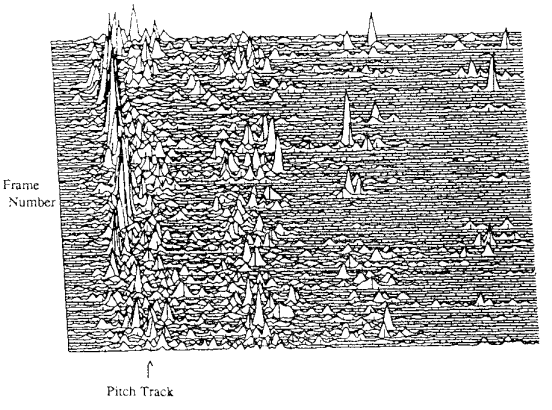


Figure 4: MUSIC-MUSIC Cepstra for Speech File with SNR = 9.5 dB

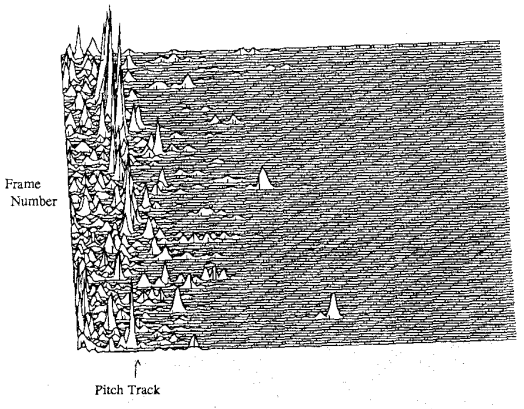


Figure 5: Distributions of Amplitudes of MUSIC Estimates

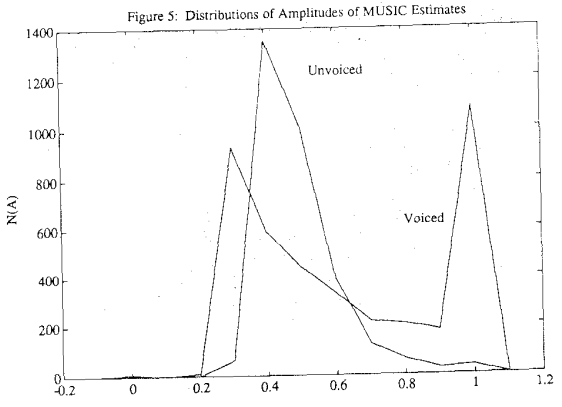


Figure 6: Singular Values of Speech Data with SNR = 5.7 dB

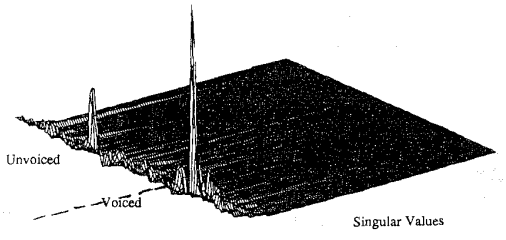


Figure 7: Singular Values of Speech Data with SNR = 18.2 dB

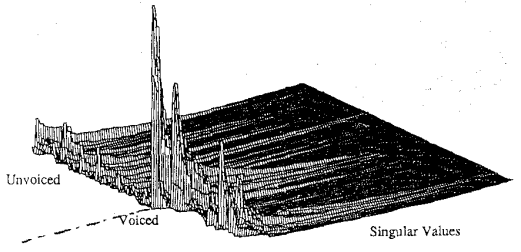


Figure 8: Singular Values of Speech Data with SNR = 30.7 dB

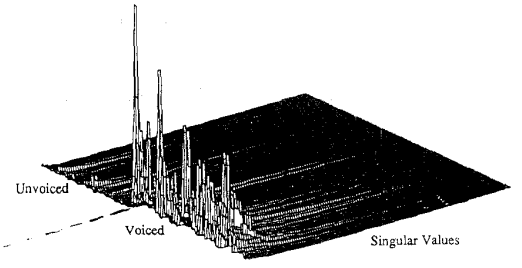


Figure 9: Distributions of Log of First Eigenvalue for Database

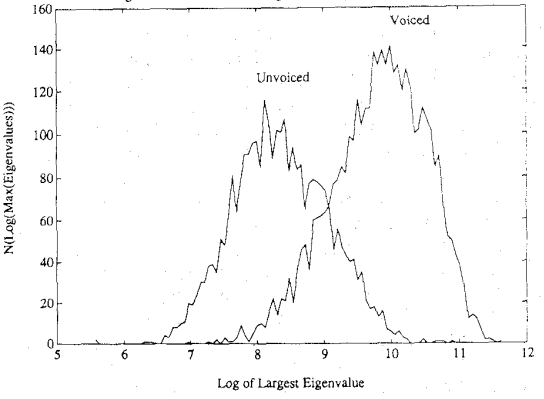
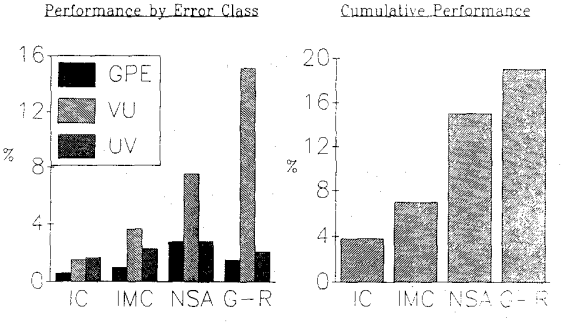


Figure 10: Performance Comparison with Leading Algorithms



IC: Integrated Correlator [1]
IMC: Integrated MUSIC-log-MUSIC Cepstrum
NSA: NSA LPC-10 Dyptrack [1]
G-R: Gold-Rabiner [1]

GPE: Gross Pitch Error
VU: Voiced to Unvoiced Error
UV: Unvoiced to Voiced Error