COMPARISON OF AUTOREGRESSIVE PARAMETER ESTIMATION ALGORITHMS FOR SPEECH PROCESSING AND RECOGNITION

Robert W. Morris, Jon A. Arrowood

Nexidia Inc.
3565 Piedmont Road NE
Atlanta, GA 30305
rmorris@nexidia.com, jarrowood@nexidia.com

Mark A. Clements

Center for Signal & Image Processing
Georgia Institute of Technology
Atlanta, Georgia 30332-0250
clements@ece.gatech.edu

ABSTRACT

Noise mitigation systems for speech coding and recognition have primarily focused on spectral subtraction techniques due to their well understood behavior and computational simplicity. As computation complexity becomes a smaller constraint, understanding the characteristics of different estimation schemes becomes more important. In this paper, the merits of two algorithms based on direct estimation of the linear prediction spectrum of a speech signal are explored. These algorithms are maximum likelihood (ML) and minimum mean square error estimation (MMSE) of the autoregressive speech spectrum. The MMSE algorithm is able to effectively improve objective quality at low SNRs while also improving the speech recognition accuracy by 20-30% on the Aurora2 test set at the cost of requiring two orders of magnitude more operations than the ML method. Because of these improvements, autoregressive based algorithms should be considered in the future for noise robust speech processing tasks.

1. INTRODUCTION

Noise mitigation has long been an important topic in all aspects of speech processing. In general, the field has been dominated by algorithms based on the spectral subtraction approach despite the wide variety of algorithms that have been proposed for improving the performance of speech recognition and compression algorithms in noisy environments. One class of algorithms is based on the direct estimation of the all-pole model of speech. These methods, which were first proposed in [1] and improved and applied in [2], iteratively filter the speech signal with an algorithm similar to expectation maximization (EM). These algorithms have not found widespread acceptance for a variety of reasons. Compared to traditional spectral subtraction algorithms, these algorithms require over an order of magnitude more computation. Secondly, direct use of the approximate MAP algorithm has been shown to produce erratic spectral dynamics, which must then be dealt with by various ad-hoc approaches [2].

Since these algorithms were presented, the availability of computational resources has increased dramatically, and implementations of these algorithms have become more practical. For this reason, maximum likelihood estimation of LPC parameters for speech coding has been recently studied in [3], where the direct estimates were found to be too erratic to be used alone. However, these estimates could be improved by using models of the spectral dynamics. In addition, MMSE estimation has been used for speech recognition [4], where the use of the autoregressive (AR) model was found to improve performance over a similar algorithm with no AR assumption.

The goal of this paper is to describe the mathematical similarities between ML and MMSE estimators of speech LPC parameters, while comparing their relative estimation performance and computational efficiency.

2. ALGORITHMS

We assume that the nth observed signal block $y_n[t] = s_n[t] + v_n[t], t = 1, \ldots, N$, is the sum of an autoregressive speech signal, $s_n[t]$, and a random Gaussian noise signal, $v_n[t]$, with power spectrum, $P_v(\omega)$. This is expressed by

$$s_n[t] = \sum_{k=1}^{p} a_n[k] s_n[t-k] + e_n[t], \quad \mathbf{e}_n \sim \mathcal{N}\left(\mathbf{0}, \sigma_s^2 \mathbf{I}\right). \quad (1)$$

2.1. ML Estimation

The EM algorithm finds the ML estimate of the linear prediction coefficients by constructing a sequence of estimates with increasing likelihood, $\theta^{[k]} = \{\hat{a}_l^{[k]}, \hat{\sigma}^{2,[k]}\}$, where k is the iteration number. The estimate $\theta^{[k]}$ maximizes the likelihood of the complete data, $\{\mathbf{s}, \mathbf{v}\}$, conditioned on the observations $\mathbf{y}$ and the previous iteration's estimate $\theta^{[k-1]}$. The algorithm is initialized by two iterations of an approximate MAP algorithm [1]. Each iteration consists of two steps: an E-step where the conditional power spectrum of the speech is calculated given the previous estimates, and an M-step where the parameters are updated. Under the assumption that the frequency components are decorrelated by the Fourier transform, the distribution for each speech spectral component is

$$S(\omega)|Y(\omega), \theta^{[k]} \sim \mathcal{N}\left(\hat{S}^{[k]}(\omega), C_S^{[k]}(\omega)\right), \quad (2)$$

where the different spectra are given by

$$\hat{S}^{[k]}(\omega) = \frac{P_s^{[k]}(\omega)}{P_s^{[k]}(\omega) + P_v(\omega)} Y(\omega), \quad (3)$$

$$C_S^{[k]}(\omega) = \frac{P_s^{[k]}(\omega) P_v(\omega)}{P_s^{[k]}(\omega) + P_v(\omega)}, \quad (4)$$

$$P_s^{[k]}(\omega) = \left| \frac{\sigma^{2,[k]}}{1 + \sum_{l=1}^{p} a_l^{[k]} e^{-jl\omega}} \right|^2. \quad (5)$$

The E-step requires the calculation of the expected squared spectral amplitude

$$\mathbb{E}\left[\,|S(\omega)|^2 \,\middle|\, Y(\omega), \theta^{[k]}\right] = \left|\hat{S}^{[k]}(\omega)\right|^2 + C_S^{[k]}(\omega). \quad (6)$$

For the M-step, we get the conditional autocorrelation sequence by taking the inverse Fourier transform

$$r_m^{[k+1]} = \frac{1}{N}\sum_{\omega_l} \mathbb{E}\left[\,|S(\omega_l)|^2 \,\middle|\, Y(\omega_l), \theta^{[k]}\right] e^{j\omega_l m}. \quad (7)$$

The auxiliary function is maximized by solving

$$\hat{\mathbf{a}}^{[k]} = -\left(\mathbf{R}_s^{[k-1]}\right)^{-1}\mathbf{r}_s^{[k-1]}, \quad (8)$$

$$\hat{\sigma}^{2,[k]} = r_0^{[k-1]} - 2\hat{\mathbf{a}}^{[k],T}\mathbf{r}_s^{[k-1]} + \hat{\mathbf{a}}^{[k],T}\mathbf{R}_s^{[k-1]}\hat{\mathbf{a}}^{[k]}, \quad (9)$$

where $\mathbf{R}_s$ and $\mathbf{r}_s$ are constructed from the sequence r_m in the same manner as in the autocorrelation method. This algorithm iterates until the change in the likelihood is sufficiently small. The final estimate of the LPC spectrum is then given by the last iteration $\hat{\mathbf{a}}^K$. Nonparametric feature vectors such as MFCCs, which is notated by $\mathcal{M}$, can be computed using the final expected spectrum:

$$\hat{\mathbf{f}} = \mathcal{M}\left(\hat{\mathbf{S}}^{[K]}\right). \quad (10)$$

Further details can be found in [3, 5].

2.2. MMSE Estimation

Another estimation approach is to find the MMSE estimate of some function of either the LPC parameters or the speech signal. Because no analytic solution exists, we perform Monte Carlo integration on samples from $\mathbf{a}|\mathbf{y}$. Although it is difficult to find this distribution, the distributions for $\mathbf{a}|(\sigma^2, \mathbf{s}, \mathbf{y})$, $\sigma^2|(\mathbf{a}, \mathbf{s}, \mathbf{y})$, and $\mathbf{s}|(\mathbf{a}, \sigma^2, \mathbf{y})$ can be derived analytically. If they are arranged in a Gibbs sampler and sampled sequentially, these values converge in distribution to the desired random variables $\mathbf{a}|\mathbf{y}$, $\sigma^2|\mathbf{y}$, and $\mathbf{s}|\mathbf{y}$. Again, the algorithm is performed in the frequency domain to achieve the sampler:

$$\mathbf{a}^{[k]}|\sigma^{2,[k-1]}, \mathbf{s}^{[k-1]}, \mathbf{y} \sim \mathcal{N}\left(\hat{\mathbf{a}}^{[k]}, C_{\mathbf{a}}^{[k]}\right), \quad (11)$$

$$\sigma^{2,[k]}|\mathbf{a}^{[k]}, \mathbf{s}^{[k-1]}, \mathbf{y} \sim \mathcal{IG}\left(\frac{L}{2}-1, \frac{L}{2}\hat{\sigma}^{2,[k]}\right), \quad (12)$$

$$S^{[k]}(\omega)|\mathbf{a}^{[k]}, \sigma^{2,[k]}, Y(\omega) \sim \mathcal{N}\left(\hat{S}^{[k]}(\omega), C_S^{[k]}(\omega)\right). \quad (13)$$

Most of the terms are the same as in the EM algorithm described above, with the exception of the following:

$$C_{\mathbf{a}}^{[k]} = \sigma^{2,[k-1]}\left(N\mathbf{R}_s^{[k-1]}\right)^{-1}, \quad (14)$$

$$r_m^{[k]} = \frac{1}{N}\sum_{\omega_l}\left|S^{[k]}(\omega_l)\right|^2 e^{j\omega_l m}. \quad (15)$$

For each block, the sampler in Equations 11-13 creates a sequence of K samples from the posterior distributions. However, the chain requires several iterations to converge to its stationary distribution. For this reason, only samples K_b through K are considered, where K_b is the number of "burn-in" samples. From these samples, the conditional expectation of any function of the speech signal or LPC polynomial can be calculated by transforming the samples and averaging. For example, the conditional mean of any transformation of the LPC parameters is given by

$$\hat{\mathbf{a}}' = \frac{1}{K-K_b}\sum_{k=K_b+1}^{K}\mathcal{F}\left(\mathbf{a}^{[k]}\right), \quad (16)$$

while the conditional mean of the MFCCs can be calculated by

$$\hat{\mathbf{f}} = \frac{1}{K-K_b}\sum_{k=K_b+1}^{K}\mathcal{M}\left(\mathbf{S}^{[k]}\right). \quad (17)$$

Further details can be found in [4, 5].

One major issue with the Gibbs sampler algorithm is that it requires a large number of iterations to converge and produce accurate estimates. The effect of the number of iterations required was tested empirically and is nearly independent of the SNR. In general, the performance is nearly optimal after 6000 samples with 500 "burn in" samples [5]. These values are used in the following experiments.

3. PERFORMANCE DIFFERENCES

3.1. Example Spectra

To illustrate the differences between the two algorithms, the results of several simulated noisy vowels are shown in Figure 1. In this example, 20 samples are taken from the process with additive white noise at 4 dB SNR. Four different estimators are then compared: standard LPC, MMSE estimation, and ML estimation. In general, the standard LPC produces very biased estimates, while the ML estimates produce nearly unbiased estimates with very high variances. The MMSE estimator outperforms both of these methods with estimates that are unbiased and with relatively small variance.

3.2. Local Maxima

If one looks carefully at the ML spectral estimates in Figure 1, spurious formants will be noticed. These typically occur when there is a small spike in the noise in a particular frequency bin. In the case of autoregressive parameter estimation in noise, the surface can have multiple local maxima, and the EM algorithm does not guarantee convergence to the global maximum. This can yield a final spectrum that includes these spurious formants.

To illustrate these problems, synthetic data is sampled from an AR(4) process with two strong resonances. This signal is then corrupted by additive white noise. From this data, the EM algorithm is used to find the maximum likelihood estimates of an AR(2) process. With three different sets of initial conditions, there are three distinctly different solutions as shown in Figure 2. This figure also shows the noisy spectrum, $|Y(\omega)|^2$, and the noise spectrum, $P_v(\omega)$. The likelihood surface, which is parameterized in the line spectrum domain, is plotted in Figure 2. To simplify the surface to two dimensions, the log-likelihood is maximized over σ^2.

In this plot, the flat spectrum is represented by the broad peak in the likelihood, while the two resonant spectra are found by picking one of the more narrow peaks. In many cases with higher order models, the local peaks may exist for narrow formants that do not actually exist. The advantage of the MMSE estimates is that these local maxima with small regions of support have a smaller effect when they are integrated over for the final estimate. Therefore,

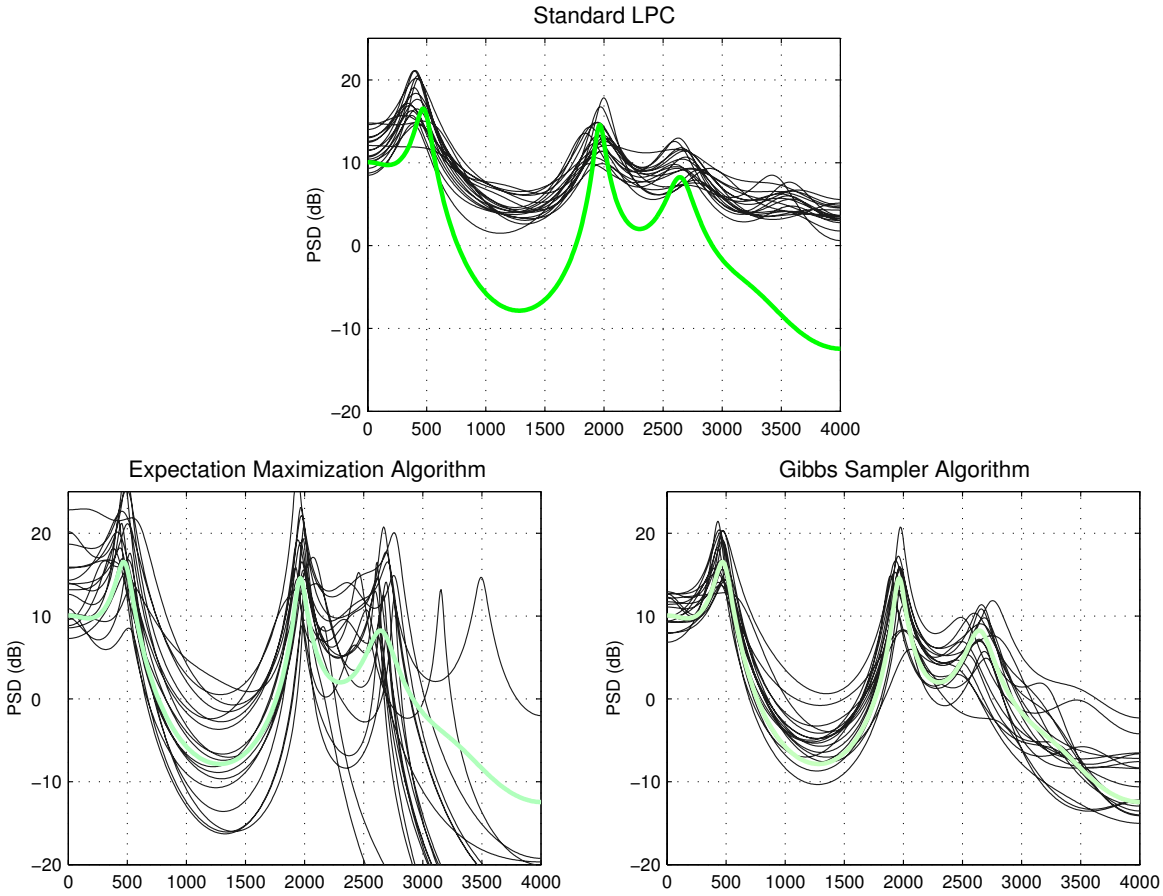

Figure 1: Spectral samples of three spectral estimation algorithms. The true spectrum is represented by the wide line with SNR = 4 dB.

this method is tends to produce fewer of these spurious formants created by the ML estimator.

3.3. Spectral Matching

To test the average performance of these estimators, noisy speech was generated by taking a whispered speech corpus and corrupting it with electronically added noise of differing SNR. The different estimation schemes were then applied to these waveforms. The resulting LPC spectra were compared to the LPC parameters of the clean waveform with the Itakura distance measure. In this paper, the median distances over the waveforms is presented.

The average spectral distortion of the ML and MMSE codecs is listed in Table 1. In all cases, the EM algorithm actually increases the spectral distance to the clean waveform, which is expected due to the erratic nature of the ML estimator. This effect was reduced in [3] by using a smoothing algorithm in the line spectrum domain based on jump Markov linear systems described in [6]. The average distortion from the smoothed estimates are also included in Table 1. The effect of the smoother on these algorithms is approximately equivalent to increasing the input SNR by 5 dB. One can also see that the performance of the MMSE algorithm is superior to the ML estimator. The distances associated with using both standard LPC and the MELPe codec, which uses a spectral subtraction type front-end enhancer, are listed at the bottom.

Table 1: Comparison of median Itakura distances for different algorithms before and after spectral smoothing.

Algorithm	**Signal to Noise Ratio**			
	0 dB	5 dB	10 dB	20 dB
ML	1.1354	0.8010	0.5777	0.0554
MMSE	0.7681	0.5654	0.3587	0.04861
ML+Smooth	0.788	0.639	0.508	0.0428
MMSE+Smooth	0.521	0.350	0.200	0.041
LPC	1.0263	0.7657	0.4778	0.1169
MELPe	0.95974	0.6765	0.4882	0.3367

3.4. Recognition Performance

These algorithms are applicable to the estimation of feature vectors for automatic speech recognition systems. The Gibbs algorithm has been tested extensively on the Aurora2 connected digits recognition task [4]. In this section, the ML and MMSE algorithms for generating feature vectors are compared using this test.

The experimental setup uses 13-dimensional MFCCs estimated using Equation 10 for ML estimation and Equation 17 for the MMSE estimator. To simplify the comparison, an ideal noise estimator averaged over 15 frames of the noise signal was used to create the noise spectral model. The models were trained using clean audio and tested over the four noise types of test set A.

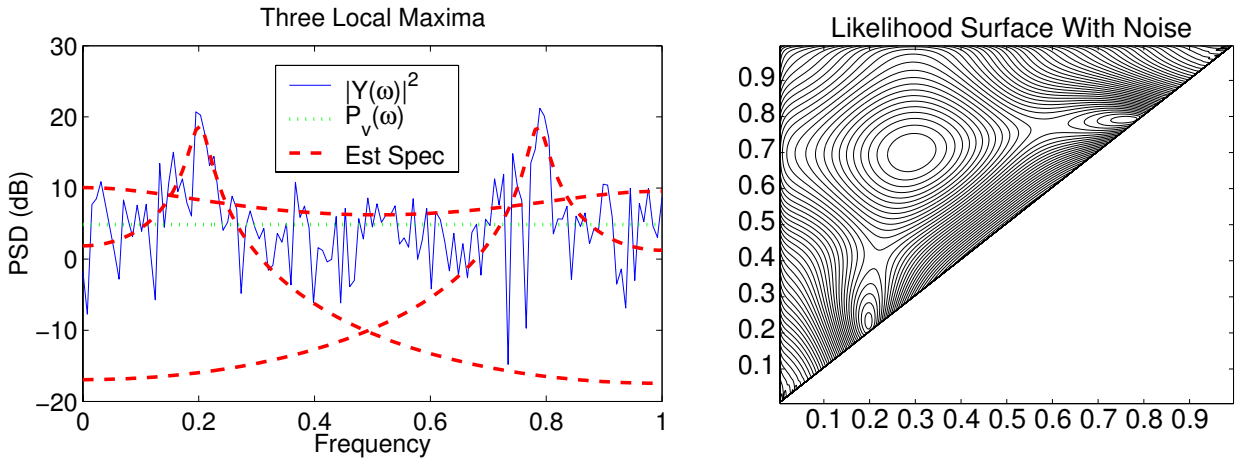

Figure 2: Example of multiple local maxima for noisy AR estimation. The left plot shows the spectrum associated with three local maxima in the likelihood surface. The right plot shows the likelihood surface. The x and y axes represent the first and second line spectral pair frequencies of the AR(2) polynomial, respectively.

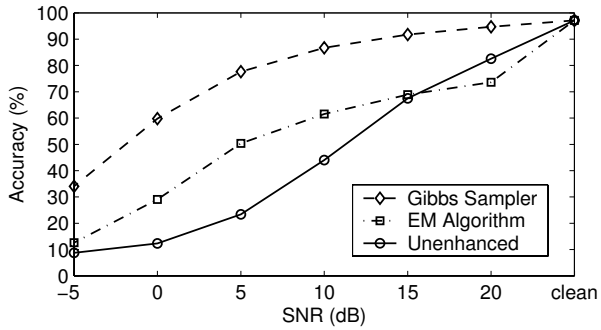

Figure 3: Comparison of ML and MMSE algorithms, averaged over set A, across SNRs.

The results are shown in Figure 3. One can see that the performance of the MMSE algorithm is superior to the ML algorithm across the SNRs. The ML algorithm performance very rapidly deteriorates at the higher SNRs, but levels off as more noise is added. This is due to the fact that the EM algorithm produces more erratic estimates in low intensity sections of the waveform. Overall, the MMSE estimates produce results that are 20-30% better than the the ML estimator.

3.5. Computation

In both methods, each iteration requires approximately the same amount of computation. In the ML estimator the expected value of the speech spectrum is taken, while it must be drawn from a distribution in the MMSE estimator . In addition to the ML calculations, the MMSE estimator requires the covariance of the LPC estimates, which requires several order p matrix operations. In addition, the Gibbs sampler requires a final averaging phase that increases the computation slightly. The major difference is in the number of iterations required. For the ML estimator, 20 to 100 iterations is sufficient for convergence, while 6000 iterations are required for the Gibbs sampler.

4. CONCLUSION

Both ML and MMSE estimation of linear prediction spectra have been described in this paper. Although they both have very similar computational structure, the stochastic MMSE estimator requires approximately one hundred times the computation of the equivalent ML estimator. However, the MMSE method in general produces more consistent results and superior performance in both spectral distance and automatic speech recognition results. In addition, the MMSE-based techniques also benefit from the flexibility gained from being able to estimate any function of the clean speech. This makes these algorithms applicable to any speech processing task included analysis, coding, and recognition.

5. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 26, pp. 197–210, June 1978.

[2] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. on Signal Processing*, vol. 39, pp. 795–805, April 1991.

[3] R. W. Morris, M. A. Clements, and J. S. Collura, "Autoregressive parameter estimation of speech in noise," in *IEEE Speech Coding Workshop*, 2002, pp. 181–183.

[4] R. W. Morris, J. A. Arrowood, and M. A. Clements, "Markov chain Monte Carlo methods for noise robust feature extraction using the autoregressive model," in *Eurospeech*, 2003, pp. 3097–3100.

[5] R. W. Morris, *Enhancement and Recognition of Whispered Speech*, Ph.D. thesis, Georgia Institute of Technology, 2003.

[6] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, pp. 515–520, 2002.