ON SECOND ORDER STATISTICS AND LINEAR ESTIMATION OF CEPSTRAL COEFFICIENTS

Yariv Ephraim

Department of ECE
George Mason University
Fairfax, VA 22030

Mazin Rahim

AT&T Labs
180 Park Avenue
Florham Park, NJ 07932

ABSTRACT

Explicit expressions for the second order statistics of cepstral components representing clean and noisy signal waveforms are derived. The noise is assumed additive to the signal, and the spectral components of each process are assumed statistically independent complex Gaussian random variables. The key result developed here is an explicit expression for the cross-covariance between the log-spectra of the clean and noisy signals. In the absence of noise, this expression is used to show that the covariance matrix of cepstral components representing a vector of N signal samples, approaches a fixed, signal independent, diagonal matrix at a rate of $1/N^2$. In addition, the cross-covariance expression is used to develop an explicit linear minimum mean square error estimator for the clean cepstral components given noisy cepstral components. Recognition results on the ten English digits using the fixed covariance and linear estimator are presented.

1. INTRODUCTION

Cepstral analysis has been widely used in signal processing. We study second order statistical properties of cepstral components and apply them to HMM-based speech recognition in clean and noisy environments. The noise is assumed additive to the signal, and the spectral components of each process are assumed statistically independent Gaussian random variables.

There are many different cepstral representations of a given signal vector. The simplest non-parametric representation of a vector $y = (y_0, \ldots, y_{N-1})^T$, with discrete Fourier transform (DFT) $Y = (Y_0, \ldots, Y_{K-1})^T$, to be studied here, is given by

$$c_y(n) = \frac{1}{K} \sum_{k=0}^{K-1} \log\left(\frac{1}{N}|Y_k|^2\right) \exp\left\{j\frac{2\pi}{K}kn\right\}, \quad (1)$$

where $K \geq 2N-1$ so that $\frac{1}{N}|Y_k|^2$ represents the DFT of linear rather than circular sample correlation of the vector y. It is shown in Section 2 that $c_y(n)$ is stable even though it is derived from the unstable periodogram estimate $\frac{1}{N}|Y_k|^2$ of the power spectral density of the signal.

In speech recognition applications, $N-$dimensional signal vectors y are represented by L-dimensional cepstral feature vectors $(c_y(n),\ n = 1, \ldots, L)^T$, where $L \ll N$. The zeroth component $c_y(0)$ is often excluded when representation of the log-spectrum average is not desired. The cepstral vectors from a given word in the vocabulary are assumed to have the probability density function (pdf) of an HMM with Gaussian state-dependent pdf's. Each state is assumed to have its own mean vector and diagonal covariance matrix. A theoretical justification for attributing diagonal covariance matrices to cepstral vectors was given in [1].

In addition to providing significant reduction in dimensionality, the cepstral representation of acoustic speech signals captures the spectral envelop of the signal while suppressing the speaker dependent pitch information, it reduces the dynamic range of the signal in a manner similar to that performed by the human auditory system, and it enables straightforward equalization of transmission channels. All of these useful properties are in fact a direct consequence of using the logarithm function [2]. Unfortunately, however, this nonlinear logarithmic function creates major difficulties when the recognizer is trained on clean speech signals but is used to recognize signals corrupted by additive noise. In this case the effect of the noise on the cepstral representation of the clean signal is rather difficult to quantify. Furthermore, suppression of the noise in the cepstral domain requires careful analysis beyond the standard Wiener filtering theory. These subjects constitute the main focus of this work.

The key result developed here is an expression for the cross-covariance between the log-spectra of the clean and noisy signals. This cross-covariance is given by an infinite weighted sum of powers of the Wiener filter for estimating the clean signal from the noisy signal.

Given this expression we first show that the covariance of cepstral components rapidly approaches a fixed signal independent constant diagonal matrix. Secondly, we derive an explicit linear minimum mean square error (MMSE) estimator for cepstral components of the clean signal from corresponding cepstral components of the noisy signal. The fixed covariance allows reduction in the number of HMM parameters by at least a factor of two. The linear MMSE estimator can be used as a preprocessor when the input signal is noisy.

2. MAIN THEORETICAL RESULTS

The derivations of the results in this section can be found in [2]. Let y, w and z denote $N-$dimensional vectors of the clean signal, the noise process and the noisy signal, respectively. The noise is additive so $z = y + w$. Let $\bar{Y}_k$, $\bar{W}_k$ and $\bar{Z}_k$, respectively, denote k-th normalized DFT components of the signal, noise, and noisy signal. The normalization is by $N^{1/2}$ so that $|\bar{Y}_k|^2$ represents power spectral density as opposed to energy spectral density. The spectral components of each process are assumed statistically independent complex Gaussian random variables with zero mean and variances given by $E\{|\bar{Y}_k|^2\} = \lambda_{Y_k}$, $E\{|\bar{W}_k|^2\} = \lambda_{W_k}$ and $E\{|\bar{Z}_k|^2\} = \lambda_{Z_k}$ for $k = 0, \ldots, K-1$. Note that the processes themselves need not be strictly Gaussian as, under certain assumptions, their spectral components become statistically independent Gaussian random variables as a result of a central limit theorem [3, Theorem 4.4.1].

2.1. Second-order statistics

The mean and variance of the kth component of the log-spectrum of the clean signal are, respectively, given by,

$$E\{\log|\bar{Y}_k|^2\} = \begin{cases} \log(\lambda_{Y_k}) - (\gamma - \log(\frac{e^2}{2})), & k = 0,\ \frac{K}{2} \\ \log(\lambda_{Y_k}) - \gamma, & k = 1, \ldots, \frac{K}{2} - 1 \end{cases} \tag{2}$$

where $\gamma = 0.57721566490$ is the Euler constant, and $e = 2.71828$ is the natural logarithm basis, and

$$\mathrm{var}(\log|\bar{Y}_k|^2) = \begin{cases} \sum_{n=1}^{\infty} \frac{n!}{(0.5)_n} \frac{1}{n^2} & k = 0,\ \frac{K}{2} \\ \sum_{n=1}^{\infty} \frac{1}{n^2} & k = 1, \ldots, \frac{K}{2} - 1 \end{cases} \tag{3}$$

where $(a)_n \triangleq 1 \cdot a \cdot (a+1) \cdot (a+2) \cdots (a+n-1)$. Furthermore, $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$. Similar expressions for the mean and variance of the log-spectrum of the noise and the noisy process hold. The appropriate expressions are obtained by replacing Y in (2) and (3) by W and Z, respectively. The covariance between the kth components of the log-spectra of the clean signal and the noisy process is given by

$$\mathrm{cov}(\log|\bar{Y}_k|^2, \log|\bar{Z}_k|^2) = \begin{cases} \sum_{n=1}^{\infty} \frac{n!}{(0.5)_n} \frac{1}{n^2} G_k^n & k = 0,\ \frac{K}{2} \\ \sum_{n=1}^{\infty} \frac{1}{n^2} G_k^n, & k = 1, \ldots, \frac{K}{2} - 1 \end{cases} \tag{4}$$

where G_k denotes the Wiener filter of the spectral component $\bar{Y}_k$ given the noisy component $\bar{Z}_k$,

$$G_k = \frac{\lambda_{Y_k}}{\lambda_{Y_k} + \lambda_{W_k}}. \tag{5}$$

Several comments are in order:

1. The variance of the kth log-spectrum component of any of the three processes is the same, and is given by the constant $\pi^2/6$ for $0 < k < K/2$. A similar result was reported in [4] where stabilization of the log-periodogram using a smoothing spline was considered. An asymptotic version of this property (as $N \to \infty$) was first shown by Brillinger [3, Corollary 5.6.3] for the smoothed periodogram power spectral density estimator of a strictly stationary process with finite moments and small dependence span. No explicit assumption that the spectral components are either Gaussian or independent was made [3].
2. In the absence of noise, $\lambda_{W_k} = 0$, $G_k = 1$, and (4) reduces to (3).
3. The covariance function given in (4) for $0 < k < K/2$ is a member of the polylogarithm functions. These functions are defined as $Li_m(z) \triangleq \sum_{n=1}^{\infty} \frac{z^n}{n^m}$, where z denotes here a complex variable. The covariance function given in (4) for $0 < k < K/2$ equals $Li_2(G_k)$ and is known as the dilogarithm function.
4. Since $0 \le G_k \le 1$, the sums in (4) do not converge slower than the sums in (3). We found that an error of less than 1% occurs if the infinite sum for $0 < k < K/2$ in (3) is truncated to only 61 terms.

The second order statistics of cepstral components can be derived from the second order statistics of the log-spectrum. The mean of $c_y(n)$ is obtained from inverse DFT of (2) and is given by

$$E\{c_y(n)\} = \frac{1}{K} \sum_{k=0}^{K-1} \log(\lambda_{Y_k}) \exp\{j \frac{2\pi}{K} kn\} + \frac{1}{K} \xi_n \tag{6}$$

where

$$\xi_n \triangleq \begin{cases} 2\log(\frac{e^2}{2}) - K\gamma & \text{if } n = 0 \\ 2\log(\frac{e^2}{2}) & \text{if } n \text{ even} \\ 0 & \text{if } n \text{ odd}. \end{cases} \tag{7}$$

Similar expressions hold for $c_w(n)$ and $c_z(n)$.

The covariance of $c_y(n)$ for $n, m = 0, \cdots, K-1$ is given by

$$\begin{aligned}&\mathrm{cov}(c_y(n), c_y(m)) \\ &= \tfrac{1}{K^2} \textstyle\sum_{k=0}^{K-1} \nu_k \mathrm{var}\{\log(|\bar{Y}_k|^2)\} \cos(\tfrac{2\pi}{K} kn) \cos(\tfrac{2\pi}{K} km) \\ &= \tfrac{1}{2} \tfrac{1}{K} (\varrho_{(n+m) \bmod K} + \varrho_{(n-m) \bmod K}),\end{aligned} \tag{8}$$

where

$$\nu_k \triangleq \begin{cases} 1 & \text{if } k = 0, \frac{K}{2} \\ 2 & \text{if } k \neq \{0, \frac{K}{2}\} \end{cases} \tag{9}$$

and for $n = 0, \cdots, K-1$,

$$\varrho_n \triangleq \frac{1}{K} \sum_{k=0}^{K-1} \nu_k \mathrm{var}(\log |\bar{Y}_k|^2) \exp\{j \frac{2\pi}{K} kn\}. \tag{10}$$

Note that ϱ_n is the inverse DFT of the variance sequence $\mathrm{var}(\log |\bar{Y}_k|^2)$ weighted by ν_k. If we define

$$\kappa_0 \triangleq \sum_{n=1}^{\infty} \frac{n!}{(0.5)_n} \frac{1}{n^2} \approx 4.5810 \tag{11}$$

$$\kappa_1 \triangleq \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \approx 1.6449 \tag{12}$$

then using (3) and assuming K is an even number we obtain,

$$\varrho_n = \begin{cases} 2\kappa_1 + \frac{2}{K}(\kappa_0 - 2\kappa_1) & \text{if } n = 0 \\ \frac{2}{K}(\kappa_0 - 2\kappa_1) & \text{if } n = 2, 4, \cdots, K-2 \\ 0 & \text{if } n = 1, 3, \cdots, K-1. \end{cases} \tag{13}$$

Hence, from (8), we have for $n = 0, \cdots, K/2$,

$$\begin{aligned}&\mathrm{var}(c_y(n)) = \mathrm{cov}(c_y(n), c_y(n)) \\ &= \begin{cases} \frac{2}{K}\kappa_1 + \frac{2}{K^2}(\kappa_0 - 2\kappa_1) & \text{if } n = 0, \frac{K}{2} \\ \frac{1}{K}\kappa_1 + \frac{2}{K^2}(\kappa_0 - 2\kappa_1) & \text{if } 0 < n < \frac{K}{2} \end{cases}\end{aligned} \tag{14}$$

and for $n, m = 0, 1, \cdots, K/2$, $n \neq m$,

$$\begin{aligned}&\mathrm{cov}(c_y(n), c_y(m)) \\ &= \begin{cases} \frac{2}{K^2}(\kappa_0 - 2\kappa_1) & \text{if } n - m = \pm 2, \pm 4, \cdots, \pm\frac{K}{2} \\ 0 & \text{otherwise.} \end{cases}\end{aligned} \tag{15}$$

This covariance matrix is "almost" Toeplitz with zero alternate off diagonals. The deviation from Toeplitz matrix is at the $(0,0)$ and $(K/2, K/2)$ elements. An example of the 5×5 upper left block of the covariance matrix corresponding to a vector of length $K = 8$ is shown below.

$$\begin{bmatrix} 0.4516 & 0 & 0.0403 & 0 & 0.0403 \\ 0 & 0.2460 & 0 & 0.0403 & 0 \\ 0.0403 & 0 & 0.2460 & 0 & 0.0403 \\ 0 & 0.0403 & 0 & 0.2460 & 0 \\ 0.0403 & 0 & 0.0403 & 0 & 0.4516 \end{bmatrix} \tag{16}$$

Note that the covariance matrix of the cepstral components is independent of the signal. The non-zero off diagonal elements of this matrix approach zero as $1/K^2$. Hence, for sufficiently large K (and hence, large N), the cepstral components are uncorrelated and their covariance matrix approaches a diagonal matrix given by

$$\mathrm{cov}(c_y(n), c_y(m)) \approx \begin{cases} \frac{1}{K}\frac{\pi^2}{3} & \text{if } n = m = 0, \frac{K}{2} \\ \frac{1}{K}\frac{\pi^2}{6} & \text{if } 0 < n = m < \frac{K}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Note that similar expressions can be obtained for the second order statistics of cepstral components derived from inverse discrete cosine transform (DCT) rather than inverse DFT.

We now argue that the cepstral components $c_y(n)$ obtained in (1) from the unstable periodogram estimate $|\bar{Y}_k|^2$ of the power spectral density of the signal are themselves stable. From (6), if we ignore the additive constant term ξ_n/K, we see that $E\{c_y(n)\}$ corresponds to cepstral components obtained from $E\{|\bar{Y}_k|^2\}$. If K (and N) is sufficiently large so that $E\{|\bar{Y}_k|^2\}$ represents the power spectral density of the signal y, then $E\{c_y(n)\}$ represent the "true" cepstral components of the process y. If $c_y(n)$ are considered estimates of these "true" cepstral components, then these estimates are consistent, since by (14), the variance of $c_y(n)$ tends to zero as $K \to \infty$. This observation is important since it shows that it is not necessary to use a consistent power spectral density estimate in order for the cepstral components to be consistent. Thus, cepstrum derived from the inconsistent periodogram spectral estimate is not necessarily worse than cepstrum derived from a stabilized smoothed periodogram.

The fact that the covariance matrix of cepstral components is a fixed signal independent matrix, which rapidly approaches a diagonal matrix, is very important in statistical modeling. If the pdf of cepstral vectors in a given HMM state is approximated by the normal pdf, as is commonly done, then only the mean vector of that pdf must be estimated from training data while the fixed, signal independent, theoretically calculated diagonal covariance of this pdf can be used. This can significantly reduce the number of parameters that need to be estimated from the training data. If the fixed diagonal matrix is used instead of being estimated from the training data, a reduction by a factor of two in the number of estimated parameters is achieved. It is demonstrated in Section 3 that using the fixed cepstral covariance (17) rather than estimating this covariance from training data, has no effect on the performance of the speech recognition system studied here.

The fixed signal independent covariance of cepstral

components in a given state also implies that when the signal is corrupted by noise, then only the mean of the cepstral vector is affected while its covariance matrix remains intact. Hence, if noisy cepstral vectors in a given HMM state are continued to be modeled as Gaussian, then only the mean vector of this pdf needs to be compensated for the noise.

2.2. Linear Cepstrum Estimator

The linear MMSE estimator of the clean cepstral vector c_y given the noisy cepstral vector c_z can be obtained from inverse DFT of the linear MMSE estimator of the log-periodogram of the clean signal. The latter is given by

$$\hat{\mathcal{L}_Y} = E\{\mathcal{L}_Y\} + \text{cov}(\mathcal{L}_Y, \mathcal{L}_Z)\text{cov}^{-1}(\mathcal{L}_Z, \mathcal{L}_Z)(\mathcal{L}_Z - E\{\mathcal{L}_Z\}) \quad (18)$$

where $\mathcal{L}_Y \triangleq (\log(|\bar{Y}_0|^2), \cdots, \log(|\bar{Y}_{K/2}|^2))^T$ denotes the vector of the log-periodogram components of the clean signal, and $\mathcal{L}_Z$ denotes the vector of log-periodogram components of the noisy signal. For this estimator, the mean vectors $E\{\mathcal{L}_Y\}$ and $E\{\mathcal{L}_Z\}$ are given by (2), and the covariance matrices $\text{cov}(\mathcal{L}_Z, \mathcal{L}_Z)$ and $\text{cov}(\mathcal{L}_Y, \mathcal{L}_Z)$ are diagonal with entries given by (3) and (4), respectively.

3. APPLICATION TO SPEECH RECOGNITION

The fixed cepstral covariance matrix (17) and the linear estimator (18) were tested in speech recognition of the ten English digits. We used isolated digits of 55 male speakers for training and another 56 male speakers for testing, all from the TIDIGITS date base. We have used a left-right HMM based system with 11 cepstral components for each vector of 200-256 samples of the speech signal sampled at 8kHz. The zeroth cepstral component was always excluded. No cepstral derivatives of any kind were used. The HMM for each digit had 10 states and 2 mixture components per state.

We first conducted recognition of clean signals, once with cepstral covariance matrices estimated from the data, as is usually done, and then by using the fixed covariance (17) for all states and mixture components. In both cases, the parameters of the Markov chain and the cepstral mean vectors for each state and mixture were estimated from the data. In both cases we have achieved the same average recognition word accuracy of 98.75%. This recognition score was obtained when the cepstral analysis (1) was applied to Hanning windowed nonoverlapping vectors of $N = 200$ samples of the clean signals, and DFT of $K = 400$ was used. Since there was no loss in performance when the fixed cepstral covariance matrix (17) was used, we continued to use this matrix for all subsequent experiments on noisy signals.

Next, we have tested the linear cepstral estimator as a preprocessor to our speech recognition system which was always trained on the clean signals. The noise was computer generated white Gaussian noise at SNR's of 10-30dB. The key to successful implementation of the linear cepstrum estimator is reliable estimation of the variances of spectral components of the clean signal and the noisy process in each analysis frame which are needed in (2) and (5). Such estimator was obtained using the window method of spectral estimation. Specifically, the variances of the spectral components of the noisy signal were obtained from

$$\hat{\lambda}_{Z_k} = \sum_{m=-(M-1)}^{M-1} w_P(m)\hat{r}_Z(m)\exp\{-j\frac{2\pi}{K}km\} \quad (19)$$

where $\hat{r}_Z(m)$ denotes the biased autocorrelation estimate obtained from a super-frame (longer frame of say T samples) which contains the current frame of the noisy signal, and $w_P(m)$ denotes the window of length $M \ll T$. The Parzen window was found particulary useful. A similar estimator was used for estimating the variances of the noise spectral components. The variances of the signal spectral components were obtained by subtracting $\hat{\lambda}_{W_k}$ from $\hat{\lambda}_{Z_k}$ and using an appropriate floor when the difference becomes non-positive.

The linear estimator was applied to non-windowed frames of $N = 256$ samples and spectral variances were estimated from super-frames of $N+2N/3$ samples with $M = 60$. The recognition results for the noisy and preprocessed cepstral components are shown below.

SNR [dB]	10	15	20	25	30
Noisy	31.07	50.45	73.75	84.64	91.25
PreProcessed	88.93	94.36	96.43	97.23	97.95

In summary, we have demonstrated for the given recognition task that using signal independent fixed diagonal covariance matrices did not affect recognition accuracy, and linear estimation of cepstral components significantly improved performance on noisy signals.

4. REFERENCES

[1] N. Merhav and C.-H. Lee, "On the asymptotic statistical behavior of empirical cepstral coefficients," IEEE Trans. SP , pp. 1990-1993, May 1993.

[2] Y. Ephraim and M. Rahim, "On second order statistics and linear estimation of cepstral coefficients," submitted for publication.

[3] D. R. Brillinger, Time Series-Data Analysis and Theory. HRW, Inc., New York, 1975.

[4] G. Wahba, "Automatic smoothing of the log periodogram," J. Am Stat. Assoc., Mar. 1980.