ROBUST SPEECH RECOGNITION USING HMM'S WITH TOEPLITZ STATE COVARIANCE MATRICES

William J. J. Roberts [1] *Yariv Ephraim* [2]

[1] Defence Science Technology Organisation, Information Technology Division,
PO Box 1500, Salisbury, 5108, Australia
[2]Department of Electrical and Computer Engineering,
George Mason University, Fairfax, VA 22030, USA

ABSTRACT

Hidden Markov modeling of speech waveforms is studied and applied to speech recognition of clean and noisy signals. Signal vectors in each state are assumed Gaussian with zero mean and a Toeplitz covariance matrix. This model allows short signal vectors and thus is useful for speech signals with rapidly changing second order statistics. It can also be straightforwardly adapted to noisy signals especially when the noise is additive and independent of the signal. Since no closed form solution exists for the maximum likelihood estimate of the Toeplitz covariance matrices, an expectation-maximization procedure was used and efficiently implemented. HMM's with Toeplitz as well as asymptotically Toeplitz (e.g., circulant, autoregressive) covariance matrices are theoretically and experimentally studied. While asymptotically all of these matrices provide similar performance, they differ significantly when the frame length is finite. Recognition results are provided for clean and noisy signals at 0-30dB SNR.

1. INTRODUCTION

Hidden Markov modeling of speech signals can either be applied to vectors of speech samples or to feature vectors of the signal. Feature vectors constitute a transformation of signal vectors. The most popular feature vector consists of low order cepstral components of the signal vector. Feature vectors reduce dimensionality by maintaining the most relevant characteristics of the signal for recognition applications. Since feature vectors constitute nonlinear transformation of signal vectors, they are difficult to analyze and compensate for input noise, especially when this noise is additive. This has been the stumbling block in introducing cepstral based speech recognition systems to real world applications.

Hidden Markov modeling of speech waveforms is simpler than that of feature vectors, and can be easily adapted to input signals which have been degraded by additive noise. In this case the signal mean in each state is zero and only the covariance of the signal in each state must be estimated during training. This estimation problem becomes manageable if structured covariance matrices are assumed. A useful class of structured covariance matrices is that of asymptotically Toeplitz matrices. This class includes Toeplitz, circulant, and autoregressive matrices. Circulant and autoregressive matrices have been successfully used in speech recognition applications (see, e.g., [1] and the references therein).

The goals of this paper are to investigate hidden Markov modeling of speech waveforms and to compare it with cepstral based hidden Markov modeling in recognition of clean and noisy signals. The noise is assumed additive and statistically independent of the signal. We primarily use Toeplitz covariance matrices since they allow relatively short signal vectors and thus are useful for speech signals with rapidly changing second order statistics. It was shown in [4] that all asymptotically Toeplitz covariance matrices result in the same asymptotic performance, as the frame length goes to infinity, if the signal is indeed a hidden Markov model (HMM) and its state-dependent covariance matrices are Toeplitz. For a finite dimension frame length, however, we demonstrate that hidden Markov modeling using Toeplitz covariance matrices outperforms hidden Markov modeling using other asymptotically Toeplitz matrices.

Maximum Likelihood (ML) estimation of Toeplitz covariance matrices for Gaussian signals has no explicit form. An elegant solution using the expectation-maximization (EM) algorithm was proposed in [2]-[3]. Here the Toeplitz matrix is embedded in a larger circulant matrix for which an explicit ML estimation exists. Each iteration of the EM algorithm consists of extending the signal vector and estimation of the circulant matrix of the extended vector. An efficient implementation of this algorithm is described in Section 2.

Implementation of hidden Markov modeling of

speech waveforms requires gain adaptation since the signal may be recorded under different gain conditions during training and recognition, and the gain contour of each signal utterance varies due to non-stationarity of speech signals. We have used the gain adaptation approach of [1] in the state level. This resulted in a non-iterative gain estimation approach when the signal is clean. The details of this approach are described in Section 3.

HMM's in this work were always trained on clean signals and tested on clean and noisy signals. Since waveform modeling is used here, HMM's for the noisy signals are obtained from the trained HMM's by adding an estimate of the noise covariance matrix to the signal covariance matrix in each state. The application of this approach to recognition of the ten English digits from the TIDIGITS database which have been contaminated by additive white noise is described in Section 4. It is demonstrated that the proposed Toeplitz-based hidden Markov modeling performs similarly to a cepstral based speech recognition system when the signal is clean. Both provide about 2% error rate for isolated digits. Furthermore, for noisy signals with signal to noise ratios of 0dB to 30dB, the error rate of the Toeplitz-based hidden Markov modeling system is only 2% higher than that obtained from a cepstral based system which has been retrained for the noisy signal.

2. HMM'S WITH TOEPLITZ COVARIANCES

The Toeplitz HMM is completely specified by the initial state probabilities π_β, the state transition probabilities $a_{\alpha\beta}$, and the state-dependent Toeplitz covariance matrices $\boldsymbol{R}_\beta$, where $\alpha, \beta = 1, \ldots, M$ and M is the number of HMM states. The ML estimation of a state Toeplitz covariance matrix, say $\boldsymbol{R}$, is obtained from

$$\max_{\boldsymbol{R} \in \mathcal{I}_R} \left\{ -\log\det \boldsymbol{R} - \mathrm{tr}\{\boldsymbol{R}^{-1}\boldsymbol{S}\} \right\} \tag{1}$$

where $\mathcal{I}_R$ is the set of all Toeplitz structured covariance matrices and $\boldsymbol{S} = \sum_{n=1}^{N} y_n y_n^{\#}/N$ represents the sample covariance of the N signal vectors $y_n \in \Re^K$ assigned to the particular state. Here $\#$ denotes conjugate transpose. We are interested in the set $\mathcal{I}_R = \mathcal{T}_K$, i.e. the set of $K \times K$ symmetric non-negative definite Toeplitz matrices and the set $\mathcal{I}_R = \mathcal{C}_K \subseteq \mathcal{T}_K$, i.e. the set of $K \times K$ symmetric non-negative definite circulant matrices.

The maximization in (1) is achieved using the EM approach [2],[3]. The main idea is to embed the Toeplitz covariance matrix in a larger circulant covariance matrix for which the explicit ML estimate exists. The ML estimate when $\mathcal{I}_R = \mathcal{C}_K$ in (1) is given by

$$\boldsymbol{R}_{kl} = \frac{1}{K}\left(\sum_{(i-j)=(k-l)} \boldsymbol{S}_{ij} + \sum_{(i-j)=K-(k-l)} \boldsymbol{S}_{ij} \right), \tag{2}$$

and will henceforth be denoted by $\boldsymbol{R} = \mathrm{circ}(\boldsymbol{S})$ [2]. The Toeplitz matrix is embedded as the $K \times K$ upper left block of an $L \times L$ circulant matrix. Such embedding requires extension of the K-dimensional data vectors to L-dimensional data vectors. The missing $L - K$ components are replaced by their conditional mean estimate given the K-dimensional data vectors and a current estimate of the $L \times L$ circulant covariance matrix. Thus, an EM iteration begins with an estimate of an $L \times L$ circulant covariance matrix, proceed with estimation of the $L - K$ missing data components, and ends with a new circulant covariance matrix estimate. The process is repeated until a stopping criterion is met.

More formally, for each signal vector $y_n \in \Re^K$ assigned to the state, let $x_n = (y_n^{\#}, \eta_n^{\#})^{\#}$ denote the L-dimensional complete statistics data vector with $\eta_n \in \Re^{L-K}$. In the EM approach, we attempt to maximize the difference in likelihood between successive estimates of $\boldsymbol{R} \in \mathcal{T}_K$. Let $\boldsymbol{R}'$ be a current estimate of $\boldsymbol{R}$. Using Jensen's inequality we have

$$\begin{aligned} &\log p(y|\boldsymbol{R}) - \log p(y|\boldsymbol{R}') \\ &= \log \int p(\eta|y,\boldsymbol{R}') \frac{p(y,\eta|\boldsymbol{R})}{p(y,\eta|\boldsymbol{R}')} d\eta \\ &\geq \int p(\eta|y,\boldsymbol{R}') \log \frac{p(y,\eta|\boldsymbol{R})}{p(y,\eta|\boldsymbol{R}')} d\eta \end{aligned} \tag{3}$$

To maximize the difference between successive estimates of $\boldsymbol{R}$, we must perform the following maximization

$$\begin{aligned} &\max_{\boldsymbol{R} \in \mathcal{T}_K} \int p(\eta|y,\boldsymbol{R}') \log p(y,\eta|\boldsymbol{R}) d\eta \\ &= \max_{\boldsymbol{R} \in \mathcal{T}_K} E\left\{ \log p(y,\eta|\boldsymbol{R}) | y, \boldsymbol{R}' \right\} \\ &= \max_{\boldsymbol{C} \in \mathcal{C}_L} E\left\{ \log p(y,\eta|\boldsymbol{C}) | y, \boldsymbol{C}' \right\} \end{aligned} \tag{4}$$

where $\boldsymbol{C}'$ represents a current estimate of the circulant matrix $\boldsymbol{C}$. The second equality in (4) holds when L is sufficiently larger than K. An upper bound on L was provided in [2], but this bound is too large for most applications. Good results were obtained here using $L = 2K$ as was done in [3].

The vectors x_n formed by concatenation of the vectors y_n and η_n have circulant covariance matrices by

construction. Using the standard assumption in the HMM literature that the vectors $\{x_n\}$ are statistically independent, we may write $\log p(y,\eta|\boldsymbol{C})$ as

$$\log p(y,\eta|\boldsymbol{C}) = -\frac{N}{2}\left(L\log 2\pi + \log\det \boldsymbol{C} + \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{S}_L)\right) \tag{5}$$

where

$$\boldsymbol{S}_L = 1/N\sum_{n=1}^{N}\left[\begin{array}{c} y_n \\ \eta_n \end{array}\right]\left[\begin{array}{cc} y_n^{\#} & \eta_n^{\#} \end{array}\right]. \tag{6}$$

Thus our problem becomes that of finding the circulant matrix achieving the following maximization

$$\begin{aligned}\max_{\boldsymbol{C}\in\mathcal{C}_L} & E\left\{-\log\det\boldsymbol{C} - \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{S}_L)|y,\boldsymbol{C}'\right\} \\ = & \max_{\boldsymbol{C}\in\mathcal{C}_L}\left\{-\log\det\boldsymbol{C} - \mathrm{tr}(\hat{\boldsymbol{S}}_L\boldsymbol{C}^{-1})\right\}\end{aligned} \tag{7}$$

where $\hat{\boldsymbol{S}}_L = E\{\boldsymbol{S}_L|y,\boldsymbol{C}'\}$. By replacing the $\eta_n^{\#}$ and $\eta_n\eta_n^{\#}$ terms in (6) by their conditional expected values and performing the sample averaging we obtain

$$\hat{\boldsymbol{S}}_L = \left[\begin{array}{cc} \boldsymbol{S} & -\boldsymbol{S}\boldsymbol{P}^{\#}\boldsymbol{Q}^{\#} \\ -\boldsymbol{Q}\boldsymbol{P}\boldsymbol{S} & \boldsymbol{Q}\boldsymbol{P}\boldsymbol{S}\boldsymbol{P}^{\#}\boldsymbol{Q}^{\#} + \boldsymbol{Q} \end{array}\right] \tag{8}$$

where the $(L-K)\times K$ matrix $\boldsymbol{P}$ and the $(L-K)\times(L-K)$ matrix $\boldsymbol{Q}$ are obtained from the following partition of the inverse of the current estimate of the circulant matrix

$$\boldsymbol{C}^{-1} = \left[\begin{array}{cc} \boldsymbol{A} & \boldsymbol{P}^{\#} \\ \boldsymbol{P} & \boldsymbol{Q}^{-1} \end{array}\right]. \tag{9}$$

The maximizing $\boldsymbol{C}$ of (7) is $\boldsymbol{C} = \mathrm{circ}(\hat{\boldsymbol{S}}_L)$. Hence, the maximizing $\boldsymbol{R}$ is the $K\times K$ upper left hand block of $\boldsymbol{C}$.

3. GAIN ADAPTATION

Accounting for the varying gain contours of speech signals is a key issue in waveform modeling for speech recognition. Here we use a gain adaptation approach in which HMM's are trained for gain normalized signals, and then used in conjunction with estimates of gain contours of the test signals [1]. Contrary to [1], we apply the gain adaptation approach at the state level. Thus, at each time instant, each speech vector y_t has associated with it M state based gain factors. This results in decoupling of gain and state estimation. Letting $g = \{g_{t,\beta} > 0\}$, where $t = 0,\ldots,T$ and $\beta = 1,\ldots,M\}$, gain adapted training aims at estimating the parameter set λ of the HMM as follows

$$\max_{\lambda}\max_{g} p(y,|g,\lambda), \tag{10}$$

where $p(y,|g,\lambda)$ denotes the pdf of the gain normalized signals $y_t/g_{t,\beta}$. Given a set of gain adapted trained HMM's $\{\lambda_m\}$, $m = 1,\cdots,M$, gain adapted recognition associates a test signal y with the $m-$th word obtained from

$$\max_{i}\max_{g} p(y|g,\lambda_i). \tag{11}$$

Here the gain contour of the test data is estimated and combined with each hypothesized HMM λ_i.

The maximization over the gain sequence in (10) and (13) can be simplified using standard HMM assumptions resulting in [4]

$$(g^*_{t,s_t})^2 = \frac{y_t^{\#}\boldsymbol{R}_{s_t}^{-1}y_t}{K}. \tag{12}$$

When noisy signals $\{z\}$ are only available for recognition, gain adapted recognition is performed by

$$\max_{i}\max_{g} p(z|g,\lambda_i). \tag{13}$$

where now $\{\lambda_i\}$ represents the parameters sets of HMM's for the noisy signals. As in the clean case this maximization can be considerable simplified but unfortunately we know of no closed form expression for the maximizing gain. This maximization may be performed using numerical procedures such as Newton's method or may be implemented using the EM algorithm. In [1] the following EM based solution was derived

$$g^2_{t,s_t} = 1/K\mathrm{tr}\left[(\boldsymbol{W}\boldsymbol{R}_w + (\boldsymbol{W}z_t)(\boldsymbol{W}z_t)^{\#})\boldsymbol{R}_{s_t}^{-1}\right] \tag{14}$$

where $\boldsymbol{W} = g'^2_{t,s_t}\boldsymbol{R}_{s_t}(g'^2_{t,s_t}\boldsymbol{R}_{s_t} + \boldsymbol{R}_w)^{-1}$ and g'_{t,s_t} is the previous estimate of the gain.

4. RESULTS AND DISCUSSION

Hidden Markov modeling of speech waveforms using Toeplitz and asymptotically Toeplitz covariance matrices was tested in recognition of clean and noisy signals and compared with a cepstral based system. Recognition of the ten English digits from the isolated word portion of the TIDIGITS database was studied. The noise was computer generated additive white noise at SNR of 0-30dB. The training set consisted of 56 male speakers and 56 female speakers, each contributing two utterances per digits. Separate HMM's for male and female speakers were trained, and each test utterance was compared against the 20 HMM's representing the models for male and female speakers. The test data consisted of 56 male and 57 female speakers which were different from those in the training set. The data was downsampled to 8Khz. Left to right HMM's with $M = 18$ states were exclusively

used with a single Gaussian pdf in each state. The dimension of the signal vectors in waveform modeling was 80 samples. For cepstral modeling we used signal vectors of 500 samples, with overlapping of 400 samples, from which 20 cepstral components were estimated. The cepstral components were estimated from smoothed periodograms. No cepstral derivatives of any kind were used.

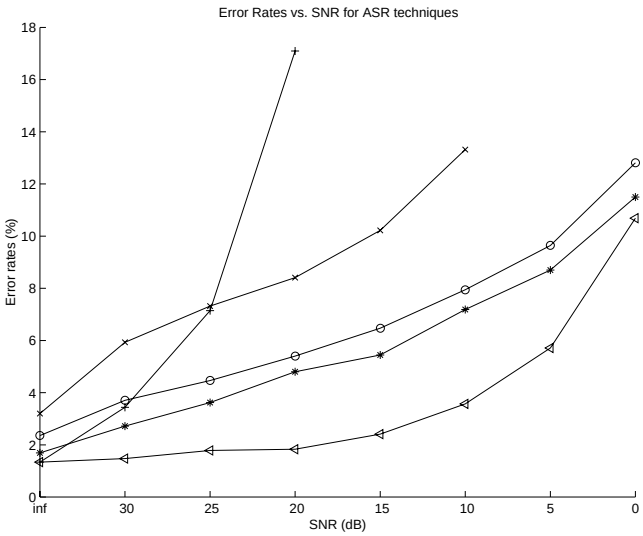

Figure 1: Performance of ASR systems in noise for: waveform HMM with Toeplitz covariances ($*$), waveform HMM with circulant covariances ($\circ$), waveform HMM with AR covariances ($\times$), cepstral HMM trained on clean signals ($+$), cepstral HMM retrained on each noise level ($\triangleleft$).

The following five speech recognition systems have been compared: 1. HMM for the signal waveforms with a Toeplitz covariance matrix in each state, 2. HMM for the signal waveforms with a circulant covariance matrix in each state, 3. HMM for the signal waveforms with an autoregressive covariance matrix of order 12 in each state, 4. HMM for cepstral vectors trained on clean signals, 5. HMM for cepstral vectors retrained on the noisy signals for each noise condition. All waveform based systems were trained on clean signals and used gain adaptation. These systems were compensated for noise by appropriate modification of their state covariance matrices.

The performance of these systems in recognizing clean and noisy speech are shown in Fig. 1. For clean signals, both the waveform HMM with Toeplitz covariance and the cepstral based HMM provided comparable results where the cepstral HMM was slightly better. In both cases the error rate was below 2%. The other waveform HMM's provided error rates higher than 2%. For noisy signals, the best performance was achieved, as expected, by the cepstral HMM which was retrained for each noise condition. When retraining is not allowed, and no compensation for noise is performed, the cepstral HMM system trained on the clean signal was rather sensitive to the input noise and provided the worst performance at SNR's smaller than 25dB. The waveform based HMM systems show gradual performance degradation as the input SNR decreased. The waveform HMM using Toeplitz covariances outperformed those using circulant and AR covariances and provided error rates of about 2% higher than the retrained cepstral based system. This performance gap was smaller at the very low and very high SNR levels.

5. COMMENTS

We have studied hidden Markov modeling of speech signal waveforms using state dependent Gaussian distributions with zero mean and Toeplitz, as well as asymptotically Toeplitz, covariance matrices. The motivation for this work was the feasible and straightforward manner in which such models can be adapted to noisy signals when the noise is additive and statistically independent of the signal. This is in contrast with the difficulties associated with adaptation of the standard non-linear cepstral based HMM's to noisy signals when the noise is additive and statistically independent of the signal (see, e.g., [5]). For HMM's trained on the clean signals and white Gaussian noise, we have achieved error rate that in higher by only 2% than that obtained using a cepstral based system which has been retrained for each noise condition. Complexity of the proposed system is low as the EM approach needs only to be applied during training of the HMM's.

6. REFERENCES

1. Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. on Speech Proc.*, vol. 40, no. 6, pp. 1303–1316, June 1992.
2. A. Dembo, C. L. Mallows, and L. A. Shepp, "Embedding non-negative definite Toeplitz matrices in non-negative definite circulant matrices, with application to covariance estimation," *IEEE Trans. on Inform. Theory*, vol. 35, no. 6, pp. 1206–1212, Dec. 1989.
3. M. I. Miller and D. L. Snyder, "The role of likelihood and entropy in incomplete-data problems. Applications to estimating point-process intensities and Toeplitz constrained covariances," *Proc. of the IEEE*, vol. 75, no. 7, pp. 892–907, July 1987.
4. W. J. J. Roberts and Y. Ephraim, "Hidden Markov modeling of speech using Toeplitz covariance matrices," *submitted for publication.*
5. M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, Sept. 1996.