RECURSIVE ESTIMATION OF TIME-VARYING ENVIRONMENTS FOR ROBUST SPEECH RECOGNITION

Yunxin Zhao, Shaojun Wang† and Kuan-Chieh Yen‡

Department of CECS, University of Missouri, Columbia, MO 65211
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213†
VerbalTek Inc., San Jose, CA 95134‡
zhao@cecs.missouri.edu swang@cs.cmu.edu kyen@verbaltek.com

ABSTRACT

An EM-type of recursive estimation algorithm is formulated in the DFT domain for joint estimation of time-varying parameters of distortion channel and additive noise from online degraded speech. Speech features are estimated from the posterior estimates of short-time speech power spectra in an on-the-fly fashion. Experiments were performed on speaker-independent continuous speech recognition using features of perceptually based linear prediction cepstral coefficients, log energy, and temporal regression coefficients. Speech data were taken from the TIMIT database and were degraded by simulated time-varying channel and noise. Experimental results showed significant improvement in recognition word accuracy due to the proposed recursive estimation as compared with the results from direct recognition using a baseline system and from performing speech feature estimation using a batch EM algorithm.

1. INTRODUCTION

There have been active research efforts on robust speech recognition in degradation environments. In general, the focus has been on stationary or slowly time-varying conditions, e.g., stationary within a speech utterance. In such cases, environment parameters are often estimated prior to system operation from a small set of adaptation data, or from online speech using batch EM algorithm in an utterance-by-utterance fashion. The estimated parameters are then used in feature estimation or model compensation for recognition of online speech.

Certain environments are known to have fast time-varying characteristics and hence the parameters need to be tracked continuously. For example, the level of additive noise may fluctuate due to movement of noise source, and the characteristics of channel may vary due to changes in signal paths. As the result, parameter estimates obtained prior to system operation are no longer relevant to subsequent speech input, and average parameter estimates obtained by batch algorithms cannot accurately represent the underlying changing environment.

This work is partially supported by the National Science Foundation under the grant NSF-EIA-9911095 and NSF-IIS-9996042.

A number of techniques have been proposed in speech research for handling nonstationary acoustic environments and three approaches that were aimed at robust speech recognition are summarized here. In the 1st approach, noise and channel were modeled by HMMs that were trained by prior measurement data of acoustics conditions and transducer characteristics. For example, a five-state HMM was used to model machine gun noise [1], and an ergotic HMM was used to model the variation of room acoustic paths due to talker movement [2]. In such cases, the problem of parameter estimation became a simpler task of identification of the underlying state sequences of the noise or channel HMMs. Mixture models were also used for capturing the on-and-off activities of multiple noise sources [3]. In the 2nd approach, evolving environments were represented by state space models of Kalman filtering and estimation were made on time-varying additive noise parameters in cepstral or log spectral domains [4,5]. The 3rd approach used sequential EM algorithm to track additive noise mean parameters in cepstral domain [6]. In addition, Bayesian recursive estimation has been proposed recently for online speaker adaptation and was shown to offer the advantages of insensitive to parameter updating interval lengths and flexible in accommodating different forms of priors [7].

In the current work, recursive estimation is proposed for tracking both channel and noise parameters in time-varying degradation environments to improve robustness of speech recognition. The algorithm is based on the frequency-domain models of speech and noise formulated previously in [8]. The recursive algorithm updates parameter estimates in a frame-by-frame fashion, and produces approximate MMSE of speech features on the fly. The technique is applicable to many commonly used speech features that are derivable from short-time speech power-spectra, for example, LPC, MFCC, PLP. A preliminary experiment was conducted to evaluate the proposed technique for improving recognition accuracy of speaker-independent continuous speech. Time-varying channel and noise were simulated to generate degraded test speech from the TIMIT database. Significant improvement in recognition word accuracy was obtained in comparison with direct recognition using the baseline system [9] and feature estimation using the frequency-domain EM algorithm [8].

2. SPEECH, NOISE AND CHANNEL MODELS

A speech degradation environment with both channel and noise

can be described by the frequency domain models defined in [8]. The system equation is $Y_n(\omega) = \Theta(\omega)X_n(\omega) + V_n(\omega)$, with $Y_n(\omega), X_n(\omega), V_n(\omega)$ denoting the short-time discrete Fourier transforms (DFT) of degraded speech, clean speech, and noise of the analysis frame *n*, and $\Theta(\omega)$ the DFT of channel's finite impulse response θ. For a size-*N* DFT, the probability density function (pdf) of the clean speech is assumed to be a mixture of Gaussian densities $f_X(\underline{X}_n(\omega);\Lambda_X) = \sum_{i=1}^{M}\alpha_i \prod_{l=0}^{N-1}\mathcal{N}(X_n(\omega_l);0,\Phi_{XX,i}(\omega_l))$, with α_i's the mixture weights and $\Phi_{XX,i}(\omega_l)$'s the class-conditional spectral variances. The parameters Λ_X can be estimated from clean training speech and hence is assumed known. The additive noise is assumed to be autoregressive Gaussian of order p (AR(p)). The pdf of degraded speech is the same as the clean speech except that the spectral variances are changed as $\Phi_{YY,i}(\omega) = |\Theta(\omega)|^2\Phi_{XX,i}(\omega) + \Phi_{VV}(\omega)$, where $\Phi_{VV}(\omega)$ denotes spectral variance of noise. Denoting the unknown parameters of the channel and noise as λ, the conditional pdf of clean speech is $f_{X|Y}(\underline{X}_n(\omega)|\underline{Y}_n(\omega);\lambda) = \sum_{i=1}^{M}\alpha_{i|\underline{Y}_n}\prod_{l=0}^{N-1}\mathcal{N}\big(X_n(\omega_l);\mu_{X|\underline{Y}_n,i}(\omega_l),\Phi_{XX|\underline{Y}_n,i}(\omega_l)\big)$. In the current work, λ is considered to be time-varying within each speech utterance.

3. RECURSIVE ESTIMATION

3.1. Basic Formulation

A recursive estimation algorithm that updates parameter per analysis frame is considered, i.e., a new parameter estimate is computed upon acquiring an observation vector $\underline{Y}_n(\omega)$. A sequence of parameter estimates $\Lambda^{(n+1)} = \left(\lambda^{(1)},\lambda^{(2)},\cdots,\lambda^{(n+1)}\right)$ is thus generated from an observation sequence $\underline{Y}_0^n(\omega) = (\underline{Y}_0(\omega),\underline{Y}_1(\omega),\cdots,\underline{Y}_n(\omega))$, starting from an initial parameter $\lambda^{(0)}$. The time-varying environment characteristic is therefore captured into the time-trajectory of the parameter sequence $\Lambda^{(n)}$.

A variety of recursive estimation algorithms exist in the literature, with trade-offs in computational complexity, convergence rate, and stability [7]. The EM-type of recursive estimator, also referred to as sequential EM, is adopted here for estimating λ to effectively handle the incomplete data problem, i.e., clean speech is missing and only degraded speech is available. Complete data vectors are defined as $\underline{Z}_n(\omega) = (\underline{Y}_n(\omega),\underline{X}_n(\omega))$, $n = 0,1,2,\cdots$, and the auxiliary objective function is defined as

$$Q^{(n)}\left(\lambda;\Lambda^{(n)}\right) = E\left[\log f_Z(\underline{Z}_0^n(\omega);\lambda)|\underline{Y}_0^n(\omega);\Lambda^{(n)}\right]$$

Based on the assumption that $\underline{Y}_n(\omega)$'s are i.i.d. and upon using the definition of $Q_k\left(\lambda;\lambda^{(k)}\right) = E\left[\log f_Z(\underline{Z}_k(\omega);\lambda)|\underline{Y}_k(\omega);\lambda^{(k)}\right]$, one has

$$Q^{(n)}\left(\lambda;\Lambda^{(n)}\right) = \sum_{k=0}^{n} Q_k\left(\lambda;\lambda^{(k)}\right)$$

A maximization of the 2nd-order Taylor series on $Q^{(n)}\left(\lambda;\Lambda^{(n)}\right)$ around $\lambda^{(n)}$ w.r.t. λ leads to a new estimate $\lambda^{(n+1)}$ as

$$\lambda^{(n+1)} = \lambda^{(n)} + I\left(\lambda^{(n)};\Lambda^{(n)}\right)^{-1} S\left(\lambda^{(n)};\Lambda^{(n)}\right) \qquad (1)$$

where $S\left(\lambda^{(n)};\Lambda^{(n)}\right)$ and $I\left(\lambda^{(n)};\Lambda^{(n)}\right)$ are the 1st and negative 2nd order derivatives of $Q^{(n)}\left(\lambda;\Lambda^{(n)}\right)$ evaluated at $\lambda = \lambda^{(n)}$, and they are referred to as score statistic and information matrix, respectively.

For recursive computation, the score statistic and information matrix are approximated as

$$\hat{S}\left(\lambda^{(n)};\Lambda^{(n)}\right) = \rho\hat{S}\left(\lambda^{(n-1)};\Lambda^{(n-1)}\right) + \frac{\partial Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\lambda}|_{\lambda=\lambda^{(n)}}$$

$$\hat{I}\left(\lambda^{(n)};\Lambda^{(n)}\right) = \rho\hat{I}\left(\lambda^{(n-1)};\Lambda^{(n-1)}\right) - \frac{\partial^2 Q_n\left(\lambda;\Lambda^{(n)}\right)}{\partial\lambda\partial\lambda^T}|_{\lambda=\lambda^{(n)}}$$

where ρ $(0 < \rho \leq 1)$ is a forgetting factor that adjusts the effect of past data in the estimation of current λ. A large ρ should be used in slowly varying conditions to fully make use of all available data, and a small ρ should be used in fast varying conditions to de-emphasize past data. As in most gradient-based algorithms, a small positive parameter ϵ can be used to adjust the step-size of adaptive estimation. As the result, a recursive estimation formula for λ is obtained as

$$\lambda^{(n+1)} = \lambda^{(n)} + \epsilon\hat{I}\left(\lambda^{(n)};\Lambda^{(n)}\right)^{-1}\hat{S}\left(\lambda^{(n)};\Lambda^{(n)}\right) \qquad (2)$$

3.2. Parameter Estimation

As in the frequency-domain EM algorithm [8], the outer-product matrix of the DFT basis vector is defined as $B(\omega_l)$, the average posterior spectral variance is defined as $\Psi_{XX|\underline{Y}_n}(\omega_l) = \sum_{i=1}^{M}\alpha_{i|\underline{Y}_n}\Phi_{XX|\underline{Y}_n,i}(\omega_l)$, and the average posterior power spectrum is defined as $G_{XX|\underline{Y}_n}(\omega_l) = \sum_{i=}^{M}\alpha_{i|\underline{Y}_n}G_{XX|\underline{Y}_n,i}(\omega_l)$. The 1st and 2nd order derivatives of $Q_n\left(\lambda;\lambda^{(n)}\right)$ are derived below for the channel and noise parameters to facilitate computing the score statistics and information matrices.

Channel

$$\frac{\partial Q_n(\lambda;\lambda^{(n)})}{\partial\theta}|_{\lambda=\lambda^{(n)}} = \sum_{l=0}^{N-1}B(\omega_l)\left\{\frac{\Psi_{XX|\underline{Y}_n}^{(n)}(\omega_l)|Y_n(\omega_l)|^2 - \Phi_{VV}^{(n)}(\omega_l)G_{XX|\underline{Y}_n}^{(n)}(\omega_l)}{\Phi_{VV}^{2(n)}(\omega_l)}\right\}_{\theta^{(n)}}$$

$$\frac{\partial^2 Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\theta\partial\theta^T}|_{\lambda=\lambda^{(n)}} = -\sum_{l=0}^{N-1}B(\omega_l)\frac{G_{XX|\underline{Y}_n}^{(n)}(\omega_l)}{\Phi_{VV}^{(n)}(\omega_l)}$$

Noise

In order to sequentially estimate the noise AR parameters, the correlation coefficients of noise, i.e., $r_{v,k},\ k = 0, 1, \cdots, p$, are first estimated using recursive estimation, and then the AR parameters are solved by the Durbin's method at each frame n. By using the relation

$$\Phi_{VV}(\omega_l) = r_{v,0} + 2\sum_{k=0}^{N-1} r_{v,k}\cos\omega_{l,k},\quad l = 0, 1, \cdots, N-1$$

the 1st and 2nd order derivatives are derived as

$$\frac{\partial Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial r_{v,k}} = \sum_{l=0}^{N-1}\frac{\partial Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}(\omega_l)}\frac{\partial\Phi_{VV}(\omega_l)}{r_{v,k}}$$

$$= \begin{cases} \sum_{l=0}^{N-1}\frac{\partial Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}(\omega_l)} & k = 0 \\ \sum_{l=0}^{N-1}\frac{\partial Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}(\omega_l)}2\cos\omega_{l,k} & k = 1, 2, \cdots, p \end{cases}$$

$$\frac{\partial^2 Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial r_{v,k}\partial r_{v,j}} = \sum_{l=0}^{N-1}\frac{\partial^2 Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}^2(\omega_l)}\frac{\partial\Phi_{VV}(\omega_l)}{r_{v,k}}\frac{\partial\Phi_{VV}(\omega_l)}{r_{v,j}}$$

$$= \begin{cases} \sum_{l=0}^{N-1}\frac{\partial^2 Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}^2(\omega_l)} & k = j = 0 \\ \sum_{l=0}^{N-1}\frac{\partial^2 Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}^2(\omega_l)}2\cos\omega_{l,k} & k = 1, \cdots, p,\ j = 0 \\ \sum_{l=0}^{N-1}\frac{\partial^2 Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}^2(\omega_l)}4\cos\omega_{l,k}\cos\omega_{l,j} & k, j = 1, \cdots, p \end{cases}$$

where

$$\frac{\partial Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}(\omega_l)}\Big|_{\lambda=\lambda^{(n)}} = -\frac{1}{2\Phi_{VV}^{(n)}(\omega_l)} + \frac{1}{2\Phi_{VV}^{2(n)}(\omega_l)}C_n(\omega_l)$$

$$\frac{\partial^2 Q_n\left(\lambda;\lambda^{(n)}\right)}{\partial\Phi_{VV}^2(\omega_l)}\Big|_{\lambda=\lambda^{(n)}} = \frac{1}{2\Phi_{VV}^{2(n)}(\omega_l)} - \frac{1}{\Phi_{VV}^{3(n)}(\omega_l)}C_n(\omega_l)$$

and

$$C_n(\omega_l) = E\left[\left|Y_n(\omega_l) - \Theta^{(n)}(\omega_l)X_n(\omega_l)\right|^2 |Y_n(\omega_l);\lambda^{(n)}\right]$$

$$= |Y_n(\omega_l)|^2 - 2Re\left(Y_n^*(\omega_l)\Theta^{(n)}(\omega_l)\mu_{X|\underline{Y}_n}^{(n)}(\omega_l)\right)$$

$$+ \left|\Theta^{(n)}(\omega_l)\right|^2 G_{XX|\underline{Y}_n}^{(n)}(\omega_l)$$

Note that a white Gaussian noise with the energy parameter σ_v^2 is a Gaussian AR(0) process with $\sigma_v^2 = r_{v,0}$.

Speech Features

The speech features of analysis frame n are estimated using $\lambda^{(n)}$ by the formula $\hat{\underline{c}}_n \simeq \sum_{i=1}^{M}\alpha_{i|\underline{Y}_n} f\left(G_{XX|\underline{Y}_n,i}^{(n)}(\omega)\right)$, which was shown to provide an approximate MMSE estimate of any speech features derivable from short-time power spectra [8]. In batch EM, $G_{XX|\underline{Y}_n,i}(\omega)$ was estimated using the converged parameter estimate λ^*; in recursive estimation, $G_{XX|\underline{Y}_n,i}^{(n)}(\omega)$ is estimated using the most recently updated $\lambda^{(n)}$, i.e., the features are estimated on the fly. The latter has the obvious advantage of alleviating time delay, and at the same time enabling feature estimation using tracked rather than averaged environment parameters. Dynamic speech features are computed via temporal regression from several successive instantaneous speech features, incurring insignificant delay.

4. EXPERIMENTS

Experiments were performed on continuous speech of the TIMIT database which were down-sampled to 10.67 KHz. The training and test speech data sets were the same as used in [8]. In the current work, the test data was degraded by simulated time-varying noise and channel. Noise samples were generated by modulating the energy of white noise with a 1–Hz cosine function over a dynamic range of 10 dB. The time-varying channel was simulated by interpolating the FIR filter parameters (50 coefficients) of a distortion channel with that of a unit impulse response, where the distortion channel was the same as used in [8] and the interpolation parameter was modulated by a 0.5–Hz cosine function. The resulting SNR in each test sentence varied between 10 dB to 20 dB. It is recognized that the current simulation of time-varying channel merely serves the purpose of testing the capability of the proposed algorithm in tracking channel variation, and more realistic time-varying channels need to be based on measurements of real world conditions.

Short-time power spectra were computed using 256–point FFTs with zero-padding and without tapering window. Speech feature vectors each consisted of 18 components: 8 PLP cepstral coefficients and log energy, and first-order temporal regression coefficients with the regression intervals of 50 ms. The speech recognizer was a speaker-independent continuous speech recognition system based on phone unit HMMs [9]. Each phone-unit HMM had three tied-states, and each state was modeled by a Gaussian mixture density. Decoding was based on time-synchronous beam search using a word-pair grammar. The recognition task had a vocabulary size of 853 and a test set perplexity of 64. Recognition word accuracy on the clean speech test set (186 sentences) was 92.6%. In recursive estimation of λ, a forgetting factor of $\rho = 0.5$ was used, and the step sizes of noise and channel were set as $\epsilon_V = 1$ and $\epsilon_\theta = 0.02$, respectively.

In Fig. 1, the behavior of the recursive estimation algorithm in tracking the time-varying noise parameter is illustrated. The parameter σ_v^2 was tracked during one speech utterance and the estimate is seen to follow the variation of the true noise parameter closely.

In order to evaluate the effect of the recursive estimation algorithm on improving speech recognition accuracy under the simulated time-varying degradation condition, speech recognition experiments were performed for the following five cases:

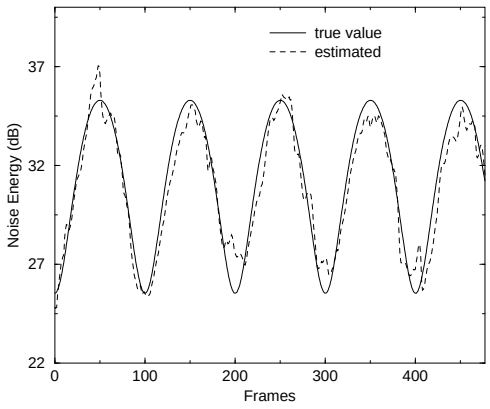

Figure 1. Estimation of the time-varying parameter σ_v^2 of white noise by the recursive estimation algorithm.

(1). Baseline — Recognized degraded speech directly;
(2). EM — Used the batch EM algorithm in [8], where the channel was initialized as a unit impulse response and noise was initialized as white.
(3). REn — Used the proposed recursive algorithm to estimate the noise parameter only, where the noise was initialized as white.
(4). RE — Used the proposed recursive algorithm, where the channel was initialized as a unit impulse response and noise was initialized as white.
(5). True-λ — Used true parameters λ in estimating the posterior speech power spectra from degraded speech.

For cases (2) through (4), the noise parameter was initialized by an estimate made over five frames of background signal immediately before the onset of each speech utterance.

Recognition word accuracies are shown in Table 1. In the case of batch EM, ten iterations were made in each utterance as if λ were time-invariant and speech features were estimated at the end of the 10th iteration. Recursive estimation RE is seen to significantly improved recognition accuracy over both baseline and EM, indicating the positive impact in tracking the time-variation of λ. RE is seen to have outperformed REn, indicating that joint estimation of noise and channel is superior to estimating noise alone. It was observed in the experiments that due to the large number of unknown FIR coefficients (50), the channel parameters were not as accurately tracked as the noise parameter as shown in Fig. 1, which could be attributed as the main factor in the performance discrepancy between RE and True-λ.

Table 1. Recognition word accuracy (%) achieved by the recursive algorithm in comparison with those of baseline, batch EM, and known environment.

Baseline	EM	REn	RE	True-λ
53.5%	69.0%	70.4%	78.4%	82.5%

5. DISCUSSION

A new technique is proposed for automatic speech recognition in time-varying environments with both distortion channel and additive noise. A recursive estimation algorithm is formulated in the frequency domain for tracking the time-varying parameters of channel and noise. Speech features are estimated on the fly from the posterior estimates of speech power spectra using the tracked channel and noise parameters. Experimental results on TIMIT speaker-independent continuous speech showed that the proposed technique led to well-tracked noise energy and significantly improved recognition accuracy, and it compared favorably over a previously proposed batch EM algorithm under the simulated time-varying condition. Further experimental evaluation is underway to investigate the performance of the proposed technique in other types of time-varying noise conditions.

REFERENCES

[1]. A. Varga and R. K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," *Proc. ICASSP*, pp. 845–848, Albuquerque, NM, Apr. 1990.

[2]. T. Takiguchi, S. Nakamura, and K. Shikano, "Speech Recognition for A Distant Moving Speaker Based on HMM Composition and Separation," *Proc. ICASSP*, pp. 1403–1406, Istanbul, Turkey, June 2000.

[3]. R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated Models of Signal and Background with Application to Speaker Identification in Noise," *IEEE Trans. on SAP*, Vol. 2, No. 2, pp. 245–258, Apr. 1994.

[4]. N. S. Kim, "IMM-based Estimation for Slowly Evolving Environment," *IEEE SPL*, Vol. 6, No. 5, June 1998.

[5]. K. Yao, B. Shi, P. Fung, and Z. Cao, "Residue Noise Compensation for Robust Speech Recognition in Nonstationary Noise," *Proc. ICASSP*, pp. 1125-1128, Istanbul, Turkey, June 2000.

[6]. N. S. Kim, D. K. Kim, and S. R. Kim, "Application of Sequential Estimation to Time-Varying Environment Compensation," *Proc. IEEE Workshop on Speech Recognition and Understanding*, pp. 389–395, Santa Barbara, CA, Dec. 1997.

[7]. S. Wang and Y. Zhao, "On-line Bayesian Speaker Adaptation Using Tree-Structured Transformation and Robust Priors," *Proc. of ICASSP*, pp. 977–980, Istanbul, Turkey, June 2000.

[8]. Y. Zhao, "Frequency-Domain Maximum Likelihood Estimation for Automatic Speech Recognition in Additive and Convolutive Noises,"*IEEE Trans. on SAP*, Vol. 8, No. 3, pp. 255–266, May 2000.

[9]. Y. Zhao, "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-sized Units," *IEEE Trans. on SAP*, Vol. 1, No. 3, pp. 345–361, July 1993.