RECURSIVE NOISE ESTIMATION USING ITERATIVE STOCHASTIC APPROXIMATION FOR STEREO-BASED ROBUST SPEECH RECOGNITION

Li Deng, Jasha Droppo, and Alex Acero

Microsoft Research,
One Microsoft Way
Redmond WA 98052, USA

ABSTRACT

We present an algorithm for recursive estimation of parameters in a mildly nonlinear model involving incomplete data. In particular, we focus on the time-varying deterministic parameters of additive noise in the nonlinear model. For the nonstationary noise that we encounter in robust speech recognition, different observation data segments correspond to different noise parameter values. Hence, recursive estimation algorithms are more desirable than batch algorithms, since they can be designed to adaptively track the changing noise parameters. One such design based on the iterative stochastic approximation algorithm in the recursive-EM framework is described in this paper. This new algorithm jointly adapts time-varying noise parameters and the auxiliary parameters introduced to linearly approximate the nonlinear model. We present stereo-based robust speech recognition results for the AURORA task, which demonstrate the effectiveness of the new algorithm compared with a more traditional, MMSE noise estimation technique under otherwise identical experimental conditions.

1. INTRODUCTION

Recently, we have successfully developed a class of front-end denoising algorithms based on the use of a limited set of stereo training data [3, 4]. The basic version of the algorithm has been called SPLICE, short for Stereo-based Piecewise Linear Compensation for Environment. For most of the noisy test speech data that have been collected and generated internally, we found that SPLICE has been highly effective [4]. These test data seem to have deviated only to a limited degree from the noisy speech in the stereo training data, due to a wide (and expensive) coverage of the noise conditions in the data design. More recently, we started applying SPLICE to the AURORA2 task, which has strongly constrained the coverage of the noise conditions in designing the stereo training data. We discovered in our earlier AURORA work that when the training set used to obtain the correction vectors in SPLICE are under very different noise environments than the enviroment for the test data, the SPLICE performance often becomes undesirably low [5]. One obvious solution to this mismatch problem is to normalize, in an instantaneous-SNR-dependent manner, the test and training environments before applying SPLICE. The effectiveness of this "noise-mean-normalized SPLICE" (NMN) has been demonstrated in our diagonostic experiments, where we used the exact noise cepstral values for the SPLICE normalization and obtained extremely high recognition performance in the AURORA2 task. This points to the crucial importance of accuracy of noise estimation in successful applications of SPLICE under serious mismatched conditions between the SPLICE's training and deployment environments.

In this paper, we describe an effective recursive noise estimation method using a weakly nonlinear model of the acoustic environment. The work is built upon some recent noise estimation work (e.g., [2, 7]), incorporating the new iterative stochastic approximation technique devised to achieve high accuracy in the Taylor series expansion of the nonlinear model. We will demonstrate the effectiveness of our new noise estimation method for the AURORA task using the NMN-SPLICE framework.

2. A NONLINEAR MODEL FOR ACOUSTIC ENVIRONMENT

The parametric model in the cepstral domain for the acoustic environment used in this work is the model described in detail in [1, 9]:

$$\mathbf{y} \approx \mathbf{h} + \mathbf{x} + \mathbf{C}\ln\big(\mathbf{I} + \exp[\mathbf{C}^{\mathrm{T}}(\mathbf{n} - \mathbf{h} - \mathbf{x})]\big), \tag{1}$$

where $\mathbf{y}$ and $\mathbf{x}$ are distorted and clean speech cepstral vectors, respectively. $\mathbf{n}$ and $\mathbf{h}$ are cepstral vectors for the additive noise and impulse response of convolutional distortion, respectively. $\mathbf{C}$ is the discrete cosine transform matrix. To simplify the notation, we define the vector function $\mathbf{g}(\cdot)$ of

$$\mathbf{g}(\mathbf{z}) = \mathbf{C}\ln\big(\mathbf{I} + \exp[\mathbf{C}^{\mathrm{T}}\mathbf{z}]\big). \tag{2}$$

The above model that relates $\mathbf{x}$, $\mathbf{y}$, $\mathbf{n}$ and $\mathbf{h}$ is mildly nonlinear. For developing the recursive estimation algorithm, we approximate this relationship by truncating the Taylor series expansion of the nonlinearity, around an iteratively updated operating point, up to the linear term. In this paper, we consider additive noise only, for which $\mathbf{h} = 0$. Let $\boldsymbol{\mu}_0^x$ and $\mathbf{n}_0$ be the operating points for the first-order Taylor series expansion of $\mathbf{y}$. We then have:

$$\begin{aligned}\mathbf{y} \;&=\; \mathbf{x} + \mathbf{g}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x) + \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)(\mathbf{x} - \boldsymbol{\mu}_0^x)\\ &\quad + [\mathbf{I} - \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)](\mathbf{n} - \mathbf{n}_0),\end{aligned} \tag{3}$$

where the gradient has the closed form of

$$\mathbf{G}(\mathbf{z}) = \mathbf{C}\mathrm{diag}\big(\frac{1}{1 + \exp[\mathbf{C}^{\mathrm{T}}\mathbf{z}]}\big)\mathbf{C}^{\mathrm{T}}.$$

3. RECURSIVE-EM ALGORITHM WITH AUXILIARY PARAMETERS

We first establish the statistical model for the clean speech cepstrum ($\mathbf{x}$ as a random vector) as a mixture of multivariate Gaussians:

$$p(\mathbf{x}) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x). \tag{4}$$

The speech frames are assumed to be independent and identically distributed. Missing data is the mixture component m, which requires the use of EM-like algorithms for ML estimation. In the work described in this paper, the noise cepstrum $\mathbf{n}$ is assumed to be a deterministic (rather than random) vector, which is time varying and is to be estimated.

Recursive noise parameter estimation is a solution to the recursive-EM optimization problem [7, 2, 8]:

$$\mathbf{n}_{t+1} = arg \max_{\mathbf{n}} Q_{t+1}(\mathbf{n}). \tag{5}$$

The objective function $Q_{t+1}(\mathbf{n})$ above is the conditional expectation

$$Q_{t+1}(\mathbf{n}) = E\big[\, ln\, p(\mathbf{y}_1^{t+1}, \mathcal{M}_1^{t+1}|\mathbf{n})|\mathbf{y}_1^{t+1}, \mathbf{n}_1^t\big], \tag{6}$$

where $\mathcal{M}_1^{t+1} = m_1, m_2, ..., m_{t+1}$ is the sequence of (hidden) mixture components in the clean speech model up to time $t+1$. The objective function above in the recursive-EM algorithm differs from the one in the conventional batch-EM — it is time indexed and the observation sequence is used up to that time.

In the E-step, the objective function Eq.6 is simplified to

$$\begin{aligned}
Q_{t+1} \quad (\mathbf{n}) &= \sum_{\mathcal{M}_1^{t+1}} p(\mathcal{M}_1^{t+1}|\mathbf{y}_1^{t+1}, \mathbf{n}_1^t) \cdot ln\, p(\mathbf{y}_1^{t+1}, \mathcal{M}_1^{t+1}|\mathbf{n}) \\
&= \sum_{\mathcal{M}_1^{t+1}} p(\mathcal{M}_1^{t+1}|\mathbf{y}_1^{t+1}, \mathbf{n}_1^t)[ln\, p(\mathbf{y}_1^{t+1}|\mathcal{M}_1^{t+1}, \mathbf{n}) + ln\, p(\mathcal{M}_1^{t+1})] \\
&= \sum_{\tau=1}^{t+1} \sum_{\mathcal{M}_1^{t+1}} p(\mathcal{M}_1^{t+1}|\mathbf{y}_1^{t+1}, \mathbf{n}_1^t)[ln\, p(\mathbf{y}_\tau|m_\tau, \mathbf{n}) + ln\, p(m_\tau)] \\
&= \sum_{\tau=1}^{t+1} \sum_{m=1}^{M} [\sum_{\mathcal{M}_1^{t+1}} p(\mathcal{M}_1^{t+1}|\mathbf{y}_1^{t+1}, \mathbf{n}_1^t)\delta_{m_\tau, m}] \\
&\qquad \cdot [ln\, p(\mathbf{y}_\tau|m_\tau, \mathbf{n}) + ln\, c_m] \\
&= \sum_{\tau=1}^{t+1} \sum_{m=1}^{M} p(m|\mathbf{y}_\tau, \mathbf{n}_{\tau-1})\, [ln\, p(\mathbf{y}_\tau \mid m, \mathbf{n}) + ln\, c_m] \\
&= \sum_{\tau=1}^{t+1} \sum_{m=1}^{M} \gamma_\tau(m) \cdot ln\, p(\mathbf{y}_\tau \mid m, \mathbf{n}) + Const.,
\end{aligned}$$

where $Const$ is a term independent of noise $\mathbf{n}$ to be estimated. The "occupancy" probability $\gamma_\tau(m) = p\big(m|\mathbf{y}_\tau, \mathbf{n}_{\tau-1}\big)$ above can be computed using the linearized version of the nonlinear acoustic environement model of Eq.3, which gives:

$$p(\mathbf{y}_\tau|m, \mathbf{n}) = \mathcal{N}\big[\mathbf{y}_\tau; \boldsymbol{\mu}_m^y(\mathbf{n}), \boldsymbol{\Sigma}_m^y\big], \tag{7}$$

where the mean for the given mixture component m is

$$\begin{aligned}
\boldsymbol{\mu}_m^y(\mathbf{n}) \;&=\; \boldsymbol{\mu}_m^x + \mathbf{g}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x) + \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)(\boldsymbol{\mu}_m^x - \boldsymbol{\mu}_0^x) \\
&+\; [\mathbf{I} - \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)](\mathbf{n} - \mathbf{n}_0),
\end{aligned} \tag{8}$$

and the covariance matrix is

$$\boldsymbol{\Sigma}_m^y = [\mathbf{I} + \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)]\boldsymbol{\Sigma}_m^x[\mathbf{I} + \mathbf{G}^T(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)]^T. \tag{9}$$

(The latter is clear after rewriting Eq.3 as $\mathbf{y} = [\mathbf{I}+\mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x]\mathbf{x}+\mathbf{d}$, where $\mathbf{d}$ is a deterministic term not affecting form of the covariance matrix.)

In Eqs.8 and 9, $\mathbf{n}_0$ (and $\boldsymbol{\mu}_0^x$) are the operating points for the Taylor series expansion, which are the auxiliary parameters to be jointly optimized with the noise parameter.

Using Eq.7 and the previously updated noise parameter, the occupancy probability $\gamma_\tau(m)$ is computed using Bayes rule in the E-step. Further, after introducing the forgetting factor ϵ, Eq.7 allows the objective function Q (ignoring constant term $Const$) to be fully expressed out as

$$\begin{aligned}
Q_{t+1}(\mathbf{n}_{t+1}) \;&=\; -\sum_{\tau=1}^{t+1} \epsilon^{t+1-\tau} \sum_{m=1}^{M} \gamma_\tau(m) \\
&\qquad \big[\mathbf{y}_\tau - \boldsymbol{\mu}_m^y(\mathbf{n}_\tau)\big]^T (\boldsymbol{\Sigma}_m^y)^{-1} \big[\mathbf{y}_\tau - \boldsymbol{\mu}_m^y(\mathbf{n}_\tau)\big] \\
&=\; \epsilon \cdot Q_t(\mathbf{n}_t) - R_{t+1}(\mathbf{n}_{t+1})
\end{aligned} \tag{10}$$

where

$$R_{t+1} = \sum_{m=1}^{M} \gamma_{t+1}(m) \big[\mathbf{y}_{t+1} - \boldsymbol{\mu}_m^y(\mathbf{n}_{t+1})\big]^T (\boldsymbol{\Sigma}_m^y)^{-1} \big[\mathbf{y}_{t+1} - \boldsymbol{\mu}_m^y(\mathbf{n}_{t+1})\big].$$

The value the forgetting factor ϵ is based on a tradeoff between the strength of noise tracking ability (ϵ close to zero) and the reliability of noise estimate (ϵ close to one).

To carry out the M-step, one can use stochastic approximation [8, 10] to sequentially update the noise parameter. Generalizing from [10] (Theorem 3; pg.264-265, where $\epsilon = 1$) and from [8] (Theorem 3.3, pg.2561), we proved that $Q_{t+1}(\mathbf{n}_{t+1})$ in recursion Eq.10 is maximized (using second-order Taylor series expansion and Newton-Raphson technique) via the following recursive form of noise parameter updating (i.e., recursive M-step):

$$\mathbf{n}_{t+1} = \mathbf{n}_t + \mathbf{K}_{t+1}^{-1}\mathbf{s}_{t+1}, \tag{11}$$

where

$$\begin{aligned}
\mathbf{s}_{t+1} \;&=\; \frac{\partial R_{t+1}}{\partial \mathbf{n}} \,|_{\mathbf{n}=\mathbf{n}_t} \\
&=\; \sum_{m=1}^{M} \gamma_{t+1}(m)[\mathbf{I} - \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)]^T (\boldsymbol{\Sigma}_m^y)^{-1} \\
&\qquad \big[\mathbf{y}_{t+1} - \boldsymbol{\mu}_m^y(\mathbf{n}_{t+1})\big],
\end{aligned} \tag{12}$$

and

$$\begin{aligned}
\mathbf{K}_{t+1} \;&=\; -\frac{\partial^2 Q_{t+1}}{\partial^2 \mathbf{n}} \,|_{\mathbf{n}=\mathbf{n}_t} \\
&=\; -\sum_{\tau=1}^{t+1} \epsilon^{t+1-\tau} \sum_{m=1}^{M} \gamma_\tau(m)[\mathbf{I} - \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)]^T \\
&\qquad (\boldsymbol{\Sigma}_m^y)^{-1}[\mathbf{I} - \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)]
\end{aligned} \tag{13}$$

In the same way as for Eq.10, we rewrite Eq.13 in a recursive form for efficient computation:

$$\mathbf{K}_{t+1} = \epsilon \cdot \mathbf{K}_t - \mathbf{L}_{t+1}, \tag{14}$$

where

$$\mathbf{L}_{t+1} = \sum_{m=1}^{M} \gamma_{t+1}(m)[\mathbf{I} - \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)]^T (\boldsymbol{\Sigma}_m^y)^{-1}[\mathbf{I} - \mathbf{G}(\mathbf{n}_0 - \boldsymbol{\mu}_0^x)]$$

4. IMPLEMENTATION USING ITERATIVE STOCHASTIC APPROXIMATION

Eq.11, Eq.12, and Eq.14 constitute a generic recursive-EM algorithm based on the general principle of stochastic approximation and on the approximate nonlinear model of acoustic environment. It sequentially estimates the noise vector for each frame, $\mathbf{n}_{t+1}$, using the information from its previous frames as well as from the current frame. In this section, we will describe practical considerations for implementing this algorithm.

In Eq.8, which is used in Eqs.12 and 13, the vectors $\mathbf{n}_0$ and $\boldsymbol{\mu}_0^x$ are the operating points for the truncated Taylor series expansion of the nonlinear environmental model, and need to be appropriately determined. For clean speech, $\mathbf{x}$, the operating points can be set naturally at a most approrpriate mean vector in the clean mixture speech model. Since we do not know in advance what mixture component a speech frame belongs to, we set the operating point $\boldsymbol{\mu}_0^x$ such that it is distributed over all mixture components $\boldsymbol{\mu}_m^x, m = 1, 2, ..., M$, with soft weights $p\big(m|\mathbf{y}_t, \mathbf{n}_t\big)$.

To determine $\mathbf{n}_0$, we assume that the noise does not change abruptly, and hence when a new frame, at $(t+1)$, of the observation is entered into the algorithm, a most reasonable noise estimate would be the estimate from the immediately preceding frame t. Therefore, we set the operating point of the Taylor series expansion for the noise at $\mathbf{n}_0 = \mathbf{n}_t$ (or a smoothed version of it) in the evaluation the $\mathbf{g}$ vector function and $\mathbf{G}$ matrix function in Eqs.12, 13, and 8.

A final consideration for improving the effectiveness of the recursive EM algorithm is based on our earlier experience that the accuracy of linear approximation to the nonlinear environment model is a key factor in speech enhancement performance [6]. Since the goal of the algorithm is to estimate the noise at the current frame at $(t+1)$ according to Eq.11, the operating point of the Taylor series expansion for noise can be iteratively updated after the estimation is completed at the same time frame $(t+1)$. (A smoothed version of the previous frame's estimate $\mathbf{n}_t$ is used to initialize this iteration.) This generalizes the stochastic approximation described in Section 3 into the new "iterative stochastic approximation", within the same recursive-EM framework.

Taking into account all the above implementation considerations, we describe the practical algorithm execution steps below. First, train and fix all parameters in the clean speech model: c_m, $\boldsymbol{\mu}_m^x$, and $\boldsymbol{\Sigma}_m^x$. Then, set $\mathbf{n}_1$ at $t = 1$ to be an average noise vector based on a crude speech-noise detector, and initialize $\mathbf{K}_0 = 0$. For each $t = 2, 3, ..., T$ in a noisy utterance $\mathbf{y}_t$, set iteration number $j = 1$ and execute the following steps sequentially:

- Step 1: Compute

$$\gamma_{t+1}^j(m) \equiv p\big(m|\mathbf{y}_{t+1}, \mathbf{n}_t^j\big) = \frac{p\big(\mathbf{y}_{t+1}|m, \mathbf{n}_t^j\big)c_m}{\sum_{m=1}^M p\big(\mathbf{y}_{t+1}|m, \mathbf{n}_t^j\big)c_m},$$

  where the likelihood $p\big(\mathbf{y}_{t+1}|m, \mathbf{n}_t^j\big)$ is from Eq.7.

- Step 2: Compute

$$\begin{aligned}\mathbf{s}_{t+1}^j \quad &= \quad \sum_{m=1}^M \gamma_{t+1}^j(m)[\mathbf{I} - \mathbf{G}(\mathbf{n}_t^j - \boldsymbol{\mu}_m^x)]^T(\boldsymbol{\Sigma}_m^y)^{-1} \\ &\quad \big[\mathbf{y}_{t+1} - \boldsymbol{\mu}_m^x - \mathbf{g}(\mathbf{n}_t^j - \boldsymbol{\mu}_m^x)\big], \quad \text{and} \quad (15)\end{aligned}$$

$$\begin{aligned}\mathbf{K}_{t+1}^j \quad &= \quad \epsilon \cdot \mathbf{K}_t^j - \sum_{m=1}^M \gamma_{t+1}^j(m)[\mathbf{I} - \mathbf{G}(\mathbf{n}_t^j - \boldsymbol{\mu}_0^x)]^T \\ &\quad (\boldsymbol{\Sigma}_m^y)^{-1}[\mathbf{I} - \mathbf{G}(\mathbf{n}_t^j - \boldsymbol{\mu}_0^x)]. \quad (16)\end{aligned}$$

- Step 3: Compute

$$\mathbf{n}_{t+1}^j = \mathbf{n}_t^j + \alpha \cdot [\mathbf{K}_{t+1}^j]^{-1}\mathbf{s}_{t+1}^j. \quad (17)$$

- Step 4: If $j < J$ (total number of iterations), then set $\mathbf{n}_t^{(j+1)} = \mathbf{n}_{t+1}^j$ and increment j++. Then continue the iteration by going to Step 1. If $j = J$, then increment t++ and start the algorithm again by re-setting $j = 1$ to process the next time frame until the end of the utterance $t = T$.

In Eq.17, α is an adjustible parameter that controls the updating rate for noise estimate. In our implemenetation, α is set to be inversely proportional to a crude estimate of the noise variance for each separate test utterance. α is also a function of J.

Several approximations have been made in the implementation of the above algorithm to significantly speed up computation. Among these approximations are: 1) a scalar-version implementation to avoid any matrix inversion; 2) approximation of $\gamma_t(m)$ to be either zero or one for each separate frame t; 3) use of Euclidean distance to determine m_0 that gives a single $\gamma_t(m_0) = 1$; and 4) use of the same m_0 for all within-frame iterations $j \leq J$.

5. NOISE-ROBUST SPEECH RECOGNITION

The recursive-EM based noise estimation algorithm described in this paper has been rigorously evaluated in the AURORA2 task [5]. Our basic denoising technique is SPLICE [3, 4], exploiting the avalability of stereo data (clean and noisy) in set-A of the database. The noise estimate is used in an enhanced, noise-mean-normalized (NMN) version of SPLICE, which effectively handles mismatched distortion conditions between set-A and set-B/C [5].

5.1. A baseline NMN-SPLICE system

A baseline noise estimation method used to evalute the new algorithm is direct computation of the traditional MMSE noise estimate by numerically carrying out the following integration:

$$\hat{\mathbf{n}} = \int \mathbf{n}p(\mathbf{n}|\mathbf{y})d\mathbf{n} = \frac{\int\int \mathbf{n}p(\mathbf{y}|\mathbf{x},\mathbf{n})p(\mathbf{x})p(\mathbf{n})\,d\mathbf{x}\,d\mathbf{n}}{\int\int p(\mathbf{y}|\mathbf{x},\mathbf{n})p(\mathbf{x})p(\mathbf{n})\,d\mathbf{x}\,d\mathbf{n}}, \quad (18)$$

where the clean-speech prior $p(\mathbf{x})$ is a pre-trained Gaussian distribution. The noise prior $p(\mathbf{n})$ in Eq.18 is also a Gaussian distribution, whose mean and variance are estimated from some noise frames in individual noisy test utterances. $p(\mathbf{y}|\mathbf{x},\mathbf{n})$ is computed using a nonlinear environmental distortion model in the log mel-spectral domain. The integrands in both the numerator and denominator of Eq.18 have been computed in closed functional forms. However, the complexity of these closed forms due to the nonlinearity in $p(\mathbf{y}|\mathbf{x},\mathbf{n})$ makes it impossible to carry out analytical integrations in Eq.18. Numerical integration is a most straightforward implementation, which we adopt for establishing a baseline NMN-SPLICE system.

5.2. Recognition results on the AURORA2 task

The numerical integration technique produces noise estimates independently for each noisy speech frame and for each mel-frequency component. The estimated noise is then used in NMN-SPLICE [5] to perform denoising for noise-robust speech recognition. The recognizer is a standard HTK system specified by the AURORA2 task. This baseline NMN-SPLICE system is used to evaluate the effectiveness of the new recursive-EM noise estimation technique under the otherwise identical experimental conditions.

Comparative recognition results are shown in Table 1 for the full AURORA2 evaluation test data. Sets A/B each consists of 1101 digit sequences for each of four noise conditions and for each of the 0dB, 5 dB, 10dB, 15dB, and 20dB SNRs. The same is for Set C except there are only two noise conditions. The recognition rates (%) in Table 1 are the average over all the noise conditions and over all the five SNRs. From Table 1, the new recursive-

Methods	Train-Mode	Set A	Set B	Set C	Overall
Numerical Integration	multicond.	88.97	87.89	87.80	88.30
	clean-only	85.33	85.75	83.74	85.18
Recursive EM	multicond.	91.49	89.16	89.62	**90.18**
	clean-only	87.82	87.09	85.08	**86.98**
No Denoising	multicond.	87.82	86.27	83.78	86.39
	clean-only	61.34	55.75	66.14	60.06
No Noise Normaliz.	multicond.	91.34	84.98	86.05	87.74
	clean-only	87.56	84.07	81.81	85.01

Table 1. Comparison of AURORA2 recognition rates (%) for the HMM systems using four different front-ends: 1) NMN-SPLICE using a baseline numerical-integration method for MMSE noise estimation; 2) NMN-SPLICE using the new recursive-EM method with iterative stochastic approximation; 3) AURORA-supplied standard MFCCs with no denoising; and 4) SPLICE with no noise normalization from training to test sets

EM method with iterative stochastic approximation performs better than the numerical integration method for noise estimation, within the same NMN-SPLICE system for cepstral enhencement. They are both significantly and consistently better than the earlier version of SPLICE with no noise normalization from training to test sets, and better than the standard MFCCs supplied by the AURORA task using no robust preprocessing to enhance speech features. The word error rate reduction using the new method is 27.9% for the multicondition training mode, and 67.4% for the clean-only mode, compared with the results with standard MFCCs with no enhancement. These results are highly statistically significant, based on a total of $1101 \times 10 \times 5 = 55050$ test utterences for each of the multicondition and clean-only training modes. In Table 2, detailed recognition rates (%) for each of the four noise conditions and for each of the SNRs in set-A using the new method for multicondition training are provided.

SNR	Subway	Babble	Car	Exhibition	Average
20 dB	98.53	98.64	98.51	98.64	98.58
15 dB	97.64	98.07	98.33	97.69	97.93
10 dB	95.98	96.37	96.84	95.65	96.21
5 dB	92.08	88.94	92.78	90.25	91.01
0 dB	78.02	65.57	76.83	74.42	73.71
Ave.	92.45	89.52	92.66	91.33	91.49

Table 2. Detailed recognition rates (%) using the new recursive-EM method with iterative stochastic approximation. Four noise conditions: Subway, Babble, Car, Exhibition-hall noises; SNRs from 0 dB to 20 dB in 5-dB increment; Set-A results with multi-condition training.

6. CONCLUSIONS

We present in this paper a recursive-EM algorithm, using a novel implementation technique of iterative stochastic approximation, for sequential estimation of nonstationary noise embedded in the speech signal. The algorithm tracks the time-varying noise while optimizing the auxiliary parameters employed to accurately approximate a nonlinear generative model for the observed noisy speech. Full speech recognition results for the AURORA task demonstrate the effectiveness of the new noise estimation algorithm in comparison with a more traditional, MMSE noise estimation method. Future work will extend the algorithm to treat the noise as time-varying random vectors and to estimate their distribution parameters. The algorithm will also be extended to include more complex speech models that incorporate dynamic features.

7. REFERENCES

[1] A. Acero, L. Deng, T. Kristjansson, and J. Zhang. "HMM adaptation using vector Taylor series for noisy speech recognition," *Proc. ICSLP*, Vol.3, 2000, pp. 869-872.

[2] M. Afify and O. Siohan. "Sequential noise estimation with optimal forgetting for robust speech recognition," *Proc. ICASSP*, Vol.1, 2001, pp. 229-232.

[3] L. Deng, A. Acero, M. Plumpe, and X.D. Huang. "Large-vocabulary speech recognition under adverse acoustic environments," *Proc. ICSLP*, Vol. 3, 2000, pp. 806-809.

[4] L. Deng, A. Acero, L. Jiang, J. Droppo, and XD Huang. "High-performance robust speech recognition using stereo training data," *Proc. ICASSP*, Vol.1, 2001, pp. 301-304.

[5] J. Droppo, L. Deng, and A. Acero. "Evaluation of the SPLICE algorithm on the Aurora2 database," *Proc. Eurospeech*, 2001.

[6] B. Frey, L. Deng, A. Acero, and T. Kristjansson. "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," *Proc. Eurospeech*, 2001.

[7] N.S. Kim. "Nonstationary environment compensation based on sequential estimation," *IEEE Sig. Proc. Letters*, Vol.5, 1998, pp. 57-60.

[8] V. Krishnamurthy and J.B. Moore. "Online estimation of hidden Markov model parameters based on the Kullback-Leibler information meature," *IEEE Trans. Sig. Proc.*, Vol.41, 1993, pp. 2557-2573.

[9] P. Moreno, B. Raj, and R. Stern. "A vector Taylor series approach for environment-independent speech recognition," *Proc. ICASSP*, Vol.1, 1996, pp. 733-736.

[10] D. M. Titterington. "Recursive parameter estimation using incomplete data," *J. Royal Stat. Soc.*, Vol.46(B), 1984, pp. 257-267.