Sequential Noise Compensation by A Sequential Kullback Proximal Algorithm

Kaisheng Yao[⋆]*, Kuldip K. Paliwal*[⋆†] *and Satoshi Nakamura*[⋆]

[⋆]ATR Spoken Language Translation Research Laboratories
2-2, Hikaridai Seika-cho, Souraku-gun, Kyoto, 619-0288, Japan
[†] School of Microelectronics Engineering, Griffith University, Australia
{kyao, kkp, nakamura}@slt.atr.co.jp

Abstract

We present a sequential noise compensation method based on the sequential Kullback proximal algorithm, which uses the Kullback-Leibler divergence as a regularization function for the maximum likelihood estimation. The method is implemented as filters. In contrast to sequential noise compensation method based on the sequential EM algorithm, the convergence rate of the method and estimation error after convergence can be adjusted by a relaxation factor β_t, where the sequential EM algorithm corresponds to the particular case of $\beta_t = 1.0$. Through experiments on parameter estimation and speech recognition in noise, we verified the efficacy of the algorithm.

1. Introduction

Noise compensation has been considered to be essential for speech recognition in real noisy environments. Among the approaches for noise compensation, model-based approach assumes an explicit model representing noise effects on speech features or models. For example, a non-linear transformation of mean vector in clean speech models can be carried out on the mean μ^l_{imj} of each mixture m at state i in clean speech models after estimation of noise parameters μ^l_{nj} in the log-spectral filter bank j [1],

$$\hat{\mu}^l_{imj} = \mu^l_{imj} + \log(1 + \exp(\mu^l_{nj} - \mu^l_{imj})) \quad (1)$$

where $1 \leq j \leq J$, and J is the total number of log-spectral filter banks. Superscript l indicates parameters are in the log-spectral domain. This function assumes that the noise variance is very small, and accordingly, only the mean of the acoustic models are transformed.

For model-based method, noise estimation and noise compensation are usually separated. That is, noise estimation is done before speech recognition, and then the noise compensation is followed to transform models or features for speech recognition. Speech recognition is the last stage for the noisy speech recognition. This is true when the noise is stationary and the noise statistics can be estimated before speech recognition. However, when the noise is time-varying in utterances, which is normal in real applications, or the noise statistics is not obtainable before speech recognition, we need to derive algorithm for noisy speech recognition that can handle the above situations.

We denote this approach as the sequential noise compensation, since the noise estimation and compensation are carried out sequentially during speech recognition. Recently, we have seen more and more research efforts in this direction. Kim *et al* proposed methods based on sequential EM algorithm [2] and the Interacting Multiple Model (IMM) method [3]. In case that noise can be modeled as two parts, with one part corresponding to stationary noise effects that can be compensated before speech recognition and the other part representing residual part of noise, Yao *et al* proposed a method based on sequential EM algorithm [1] and a method employing a set of extended Kalman filters [4].

One key point in the above approaches is the assumption that noise is not stationary. If the noise statistics can be obtained before speech recognition and the noise is stationary, the methods, e.g., Log-Add [1], shall be the choice for noisy speech recognition. We argue that the gain of sequential noise compensation is apparent when the stationary assumption or obtainable noise statistics can not be hold.

Hopefully, sequential noise compensation can adapt a speech recognition system to changing environments. As is well known, convergence rate and estimation error after convergence are two factors reflecting success of adaptive systems. The algorithm has to be carefully adjusted to make a performance balance between the two factors. Our recent work on a sequential Kullback proximal algorithm [5], which is an extension of the sequential EM algorithm [6], has shown that, the convergence rate and estimation error after convergence can be adjusted by choosing β_t, a relaxation factor in the algorithm. We have shown that, a smaller β_t than 1.0 can make the algorithm have faster convergence rate than that of the sequential EM algorithm, but the estimation error after convergence might be larger than the sequential EM algorithm. When applied to noise compensation, the derived method is an extension of the Log-Add noise compensation method [1], which is shown in Equation (1), to situations where noise is supposed to be time-varying.

In this paper, we give some further theoretical results of the algorithm on its asymptotic convergence property. Besides the theoretical results, we provide some experimental results to verify its efficacy.

2. Background

During speech recognition, joint likelihood of observation sequence $Y_t = [y_1 \cdots y_t]$ and state sequence $\mathbf{q}$ is propagated via Viterbi approximation, which is given as,

$$\alpha_t(i;\theta) = \alpha_{t-1}(\zeta^\star;\theta) a_{\zeta^\star i} b_i(y_t) \quad (2)$$

where subscript t denote frame index. $\zeta^\star = \arg\max_\zeta \alpha_{t-1}(\zeta;\theta) a_{\zeta i}$. $b_i(y_t)$ is the emission probability of y_t at state i given parameter θ. Sequential parameter estimation during speech recognition is carried out via Maximum Likelihood (ML) estimation, i.e.,

$$\theta(t) = \arg\max_{\theta \in R^J} l_t(\theta)$$

where $l_t(\theta)$ is the log-likelihood till frame t.

We denote $\Theta(t-1)$ as the sequentially estimated parameter $[\theta(1)\cdots\theta(t-1)]$. Due to hidden state problem in HMM, the sequential Expectation Maximization (EM) algorithm is a popular algorithm [6] to find the ML estimation of $\theta(t)$.

Alternatively, ML estimation can also be addressed by the sequential Kullback proximal algorithm [7].

Proposition 1 *The sequential EM algorithm is equivalent to the following recursion with $\beta_t = 1, t = 1, 2, ...,$*

$$\theta(t) = \arg\max_{\theta\in R^J}\{l_t(\theta) - \beta_t I_y(\theta, \Theta(t-1))\} \quad (3)$$

where $I_y(\theta, \Theta(t-1)) = \sum_q \log\frac{f(\boldsymbol{q}|Y_t;\Theta(t-1))}{f(\boldsymbol{q}|Y_t;\theta)} f(\boldsymbol{q}|Y_t;\Theta(t-1))$ is the Kullback-Leibler (K-L) divergence till frame t.

Note that our proposition is a sequential version of the batch version in Proposition 1 in [8]. The proposition can be similarly proved as in [8].

The KL divergence between successive iterates of the state posterior densities works as a regularization factor. Following the terminology in [8], we denote the above updating procedure as the sequential Kullback proximal algorithm.

3. Sequential noise compensation by the sequential Kullback proximal algorithm

The proposition 1 gives the following sequential algorithm:

Let $\theta(0)$ be the initial parameter estimate. Define $Q_t(\Theta(t-1),\theta)$ as the conditional expectation of the log likelihood of the complete data $\mathbf{q}$, Y_t until frame t given the observation and $\Theta(t-1)$. Then given y_t, the recursive update of $\theta(t)$ is given as [7],

$$\theta(t) = \theta(t-1) - \frac{\frac{\partial Q_t(\Theta(t-1);\theta)}{\partial\theta}|_{\theta=\theta(t-1)}}{\beta_t\frac{\partial^2 Q_t(\Theta(t-1),\theta)}{\partial\theta^2}|_{\theta=\theta(t-1)} + (1-\beta_t)\frac{\partial^2 l_t(\theta)}{\partial\theta^2}|_{\theta=\theta(t-1)}} \quad (4)$$

where

$$\frac{Q_t(\Theta(t-1),\theta)}{\partial\theta}|_{\theta=\theta(t-1)} = \sum_{i=1}^{N}\gamma_t(i;\theta(t-1))\frac{\partial\log b_i(y_t)}{\partial\theta}|_{\theta=\theta(t-1)} \quad (5)$$

$$\frac{\partial^2 Q_t(\Theta(t-1),\theta)}{\partial\theta^2}|_{\theta=\theta(t-1)} = \frac{\partial^2 Q_{t-1}(\Theta(t-2),\theta)}{\partial\theta^2}|_{\theta=\theta(t-1)} + \sum_{i=1}^{N}\gamma_t(i;\theta(t-1))\frac{\partial^2\log b_i(y_t)}{\partial\theta^2}|_{\theta=\theta(t-1)} \quad (6)$$

where $\gamma_t(i;\theta)$ is the posterior probability at state i given observation sequence Y_t and parameter θ.

Exact calculation of the second order differentiation of $l_t(\theta)$ is computationally expensive. We propose an approximation, in which the value is approximated at θ around $\theta(t-1)$ as [7],

$$\frac{\partial^2 l_t(\theta)}{\partial\theta^2} = \sum_{i=1}^{N}\sum_{m=1}^{M}\gamma_t(i,m;\theta)[(\frac{\partial\log b_{im}(y_t)}{\partial\theta})^2 + \frac{\partial^2\log b_{im}(y_t)}{\partial\theta^2}] - (\sum_{i=1}^{N}\sum_{m=1}^{M}\gamma_t(i,m;\theta)\frac{\partial\log b_{im}(t)}{\partial\theta})^2 \quad (7)$$

where N and M each denote the number of states and the number of mixtures in each state. $b_{im}(y_t)$ is the emission probability of y_t at state i and mixture m given parameter $\theta(t)$. c_{im} is the Gaussian mixture weight, and $\sum_{m=1}^{M} c_{im} = 1$. $b_i(y_t) = \sum_{m=1}^{M} c_{im} b_{im}(y_t)$ and $\gamma_t(i,m;\theta) = \gamma_t(i;\theta)\frac{c_{im}b_{im}(y_t)}{b_i(y_t)}$.

In particular, for $\theta(t) = \mu^l_{nj}(t)$ representing estimated time-varying noise parameter, the noise parameter updating requires the following calculations in each mixture m at state i.

$$\frac{\partial\log b_{im}(y_t;\theta)}{\partial\theta} = \sum_{k=1}^{K}[d_{kj}\frac{(y_t(k)-\mu_{imk}(t-1))}{\sigma^2_{imk}}\frac{\partial\mu^l_{imj}(t)}{\partial\theta}]$$

$$\frac{\partial^2\log b_{im}(y_t;\theta)}{\partial\theta^2} = \sum_{k=1}^{K}[-\frac{1}{\sigma^2_{imk}}d^2_{kj}(\frac{\partial\mu^l_{imj}(t)}{\partial\theta})^2 + \frac{y_t(k)-\mu_{imk}}{\sigma^2_{imk}}d_{kj}\frac{\partial^2\mu^l_{imj}(t)}{\partial\theta^2}]$$

where $\frac{\partial\mu^l_{imj}(t)}{\partial\theta}$ and $\frac{\partial^2\mu^l_{imj}(t)}{\partial\theta^2}$ are given as $\frac{\exp(\mu^l_{nj}(t)-\mu^l_{imj})}{1+\exp(\mu^l_{nj}(t)-\mu^l_{imj})}$ and $\frac{1}{(1+\exp(\mu^l_{nj}(t)-\mu^l_{imj}))^2}$, respectively. $\mu_{imk}(t-1)$ is the corresponding compensated cepstral mean at cepstral index k, obtained after the Discrete Cosine Transform (DCT) of $\{\mu^l_{imj}(t-1) : 1\le j\le J\}$. σ^2_{imk} is the diagonal variance at cepstral index k in mixture m at state i. $y_t(k)$ is the cepstral observation element at cepstral index k in the observation vector y_t. d_{kj} is the DCT coefficient.

In fact, the algorithm works in a vector form for noise parameter updating. For simplicity in expression, we adopt the scalar formula.

3.1. Asymptotic convergence property of the algorithm

When θ approaches its maximum likelihood estimation θ_{ML}, the asymptotic convergence property of the algorithm can be approximately analyzed [7]. The analysis exploits the following observation that, when $\theta(t)$ approaches to θ_{ML}, it holds

$$E[\frac{y_t(k)-\mu_{imk}(t-1)}{\sigma^2_{imk}}|y_t;\theta(t-1)] \approx 0 \quad (8)$$

where expectation is over all states and mixtures. Following main result is that the updating by the sequential Kullback proximal algorithm can be viewed as,

$$\theta(t) \approx \theta(t-1) - \frac{\frac{\partial Q_t(\Theta(t-1);\theta)}{\partial\theta}|_{\theta=\theta(t-1)}}{\beta_t\frac{\partial^2 Q_t(\Theta(t-1),\theta)}{\partial\theta^2}|_{\theta=\theta(t-1)}} \quad (9)$$

and, accordingly, when $\theta(t)\to\theta_{ML}$, the sequential Kullback proximal algorithm keeps

$$Q_t(\Theta(t-1);\theta(t)) \ge Q_t(\Theta(t-1);\theta(t-1))$$

It is well know in the proof of the EM algorithm [9] that this makes the estimation convergent.

It is thus straightforward to compare the estimation error and convergence rate between the sequential Kullback proximal algorithm and the sequential EM algorithm. Denote the estimation by the sequential EM as θ_{SEM}, and the estimation by the sequential Kullback proximal algorithm as θ_{SKP}, by Equation (9), the ratio of the estimation error after convergence is given as,

$$\lim_{\theta \to \theta_{ML}} \frac{\|\theta_{SEM} - \theta_{ML}\|^2}{\|\theta_{SKP} - \theta_{ML}\|^2} \approx \beta_t^2 \qquad (10)$$

It shows that the sequential EM has smaller estimation error compared to the sequential Kullback proximal algorithm with $\beta_t \leq 1.0$.

The following formula shows the ratio of the convergence rate between the sequential EM algorithm and the sequential Kullback proximal algorithm when the parameter estimation is close to θ_{ML}.

$$\lim_{\theta \to \theta_{ML}} \frac{\frac{\|\theta_{SEM}(t) - \theta_{ML}\|}{\|\theta_{SEM}(t-1) - \theta_{ML}\|}}{\frac{\|\theta_{SKP}(t) - \theta_{ML}\|}{\|\theta_{SKP}(t-1) - \theta_{ML}\|}} \approx \beta_t \qquad (11)$$

Note that we have assumed that $\theta_{SEM}(t-1) = \theta_{SKP}(t-1)$.

From above analysis, we see that the sequential Kullback proximal algorithm has faster convergence rate than the sequential EM algorithm, when $\beta_t \leq 1.0$. But conversely, the sequential EM algorithm has β_t^2 times of the estimation error as that of the sequential Kullback proximal algorithm, when they are operating after convergence. We need to make a balance between the convergence rate and the estimation error, depending on particular requirement in system performance.

3.2. Remarks

Similar to [6], the exponential forgetting can be adopted, where the forgetting factor $0 < \rho \leq 1.0$. In the following experiments, the forgetting factor is set to 0.995.

4. Experiments

4.1. Experiments on the convergence rate of the algorithm

This section verifies that the method can have faster convergence rate than the sequential EM algorithm when applied to sequential noise parameter estimation. Experiments were performed on the TI-DIGITS database, which was down-sampled to 16kHz. Five hundred clean speech utterances from 15 speakers were used for training HMMs. Digits and silence were respectively modeled by 10-state and 3-state (including a non-emitting initial and final state) whole word HMMs with 4 diagonal Gaussian mixtures in each state. Contaminated speech was generated by artificially adding White noise to the clean speech.

Twenty-six filters were used in the binning stage, i.e., $J = 26$. The features were MFCC + C0 + Δ MFCC + Δ C0. The MFCCs were generated from the power of the Fourier transform in the binning process. The baseline system had a 1.3% Word Error Rate (WER) under clean conditions when tested by 100 utterances unseen in the training set.

To show that the sequential Kullback proximal algorithm can have faster convergence rate, we concatenate a single utterance, *15ooa.wav*, in *jr* set of TI-DIGITS database in 43 times. The utterance was contaminated by 16.0dB White noise. The sequential algorithms were initialized at 8.8dB SNR White noise. β_t is set to be 10.0, 1.0, 0.9 and 0.5. $\frac{\partial^2 Q_t(\Theta(t-1),\theta)}{\partial \theta^2}$ was initialized to be the minus 100 times of the variance of noise at each log-spectral filter bank. $\frac{\partial Q_t(\Theta(t-1),\theta)}{\partial \theta}$ was initialized to zero. The noise statistics were estimated from 5 seconds of available noise segments before the recognition process.

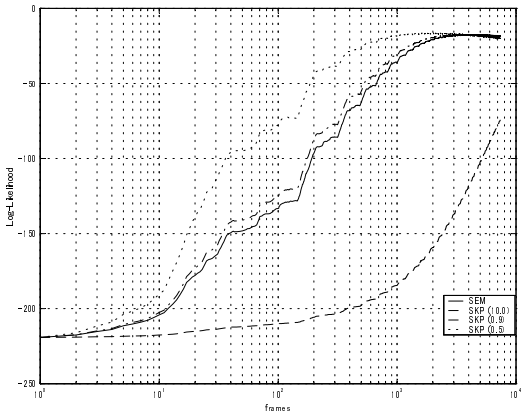

Figure 1: Plot of the Log-Likelihood versus frame index of the sequential noise compensation. Horizontal axis is the frame index in logarithm scale. Each curve (from top to bottom) represents the Log-Likelihood of the estimation at β_t equal to 0.5, 0.9, 1.0, and 10.0. Parameter updating is initialized at 8.0 dB and tested in 16.0 dB White noise.

Figure 1 shows that the Log-Likelihood of the estimation increases along with the frame index. The rate of the increase shows the rate at which the systems adapt. It shows that, the sequential Kullback proximal algorithm with $\beta_t < 1.0$ has faster adaptation rate than that by the sequential EM algorithm, while the sequential algorithm with $\beta_t > 1.0$ has slower adaptation rate than that by the sequential EM algorithm.

4.2. Experiments on speech recognition in noise

Experiments were performed on the Aurora-2 task [10]. The task uses the TI-DIGITS database downsampled from the original sampling rate of 20kHz to 8 kHz. The training set in our experiments is the clean training set, consisting of 8840 utterances. Testing utterances in our experiments are the utterances contaminated by 5dB restaurant noise and airport noise in set B of the testing set. They each labeled as N1_SNR5 and N3_SNR5 in Mfc08TS_setb of the Aurora-2 database.

The recognition system is our system, but with the HMM trained by the script provided by the Aurora-2 task. The HMMs are eleven whole word models, each comprising of 16 states with 3 mixtures in each state. Two silence models, one with 3 states and 3 mixtures to model the utterance beginning and end silence, and the other with 1 state and 6 mixtures to model the interword silence have also been used.

The feature extraction in our experiment has some differences from that provided by the Aurora-2 task. First, we use MFCC plus C0 as the static coefficients. Second, the MFCCs are generated from the power of the Fourier transform in the binning process. Last, the number of filter-banks is 26 instead of 23 used in the Aurora-2 task.

Initialization procedure was the same as that in Section 4.1, except the initial noise estimation was carried out in a different way: Before recognition of this set of utterances, we picked

up one utterance in it, and ran noise estimation by letting it go through a training network of "bk" + "gs" + "bk". The "bk" model is a one state one mixture HMM, and the "gs" model is with the same topology as the whole digit model in the Aurora-2 task. After several training iterations, the Gaussian mixture in the "bk" model will represent the statistics at the beginning and end frames of the utterance. The estimated statistics was used to initialize the sequential noise compensation procedure.

4.2.1. Experiment results

Experiment results on sequential noise compensation are shown in Table 1. Baseline denotes system with noise compensated by the sequential EM algorithm. This corresponds to setting $\beta_t = 1.0$[1]. If compensated by the Log-Add noise compensation method without sequential noise compensation, the system has 86.48% word accuracy in the restaurant noise case and 60.73% word accuracy in the airport noise case.

Table 1: Word Accuracy (in %) of the system tested in 5dB Restaurant noise and 5dB airport noise in testing set B of the Aurora-2 task. Baseline is the system with sequential noise compensation by the sequential EM algorithm. 0.5, 0.9, 2.0 and 10.0 each denote the sequential Kullback proximal algorithm with β_t set to 0.5, 0.9, 2.0 and 10.0.

	β_t				
test condition	Baseline	0.5	0.9	2.0	10.0
restaurant	78.95	78.42	78.89	79.86	81.19
airport	77.70	76.77	77.54	78.23	78.67

From Table 1, we have the following observations. First, the airport noise looks more nonstationary than the restaurant noise, as shown in Figure 2. Because of this, the initial noise estimation performance is not so reliable in the airport noise case as that in the restaurant noise case. As a result, the Log-Add noise compensation without sequential noise compensation has quite different performances in the two noises. In contrast, the sequential noise compensation procedure performs stably in the two noises. Second, although both sequential EM algorithm and sequential Kullback proximal algorithm adapt HMMs according to their estimate of noise parameters, since the larger parts of the noises are comparatively stationary, the smoother estimation (smaller estimation error) shows more influences on system performances than the faster adaptation (larger estimation error). This can be seen that the sequential Kullback proximal algorithm with $\beta_t < 1.0$ has lower word accuracy than the sequential EM algorithm, whereas it has higher word accuracy than the sequential EM algorithm when $\beta_t > 1.0$.

5. Conclusions

We have presented a sequential noise compensation method for speech recognition in time-varying noise. The method can have faster convergence rate than the sequential EM algorithm, and the estimation error after convergence can be controlled by a relaxation factor, β_t. Experiments carried out so far have verified its efficacy. We will report elsewhere further results on the method.

[1]Since our objective of this paper is to compare the performance between the sequential EM and the sequential Kullback proximal algorithm, we set the baseline to the sequential EM algorithm.

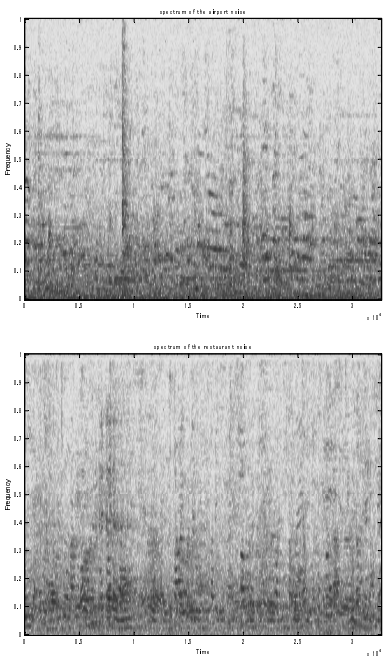

Figure 2: Spectrogram of the airport noise (upper) and the restaurant noise (lower).

6. References

[1] K. Yao, B. E. Shi, S. Nakamura, and Z. Cao, "Residual noise compensation by a sequential em algorithm for robust speech recognition in nonstationary noise," in *ICSLP*, 2000, vol. 1, pp. 770–773.

[2] N.S. Kim, "Nonstationary environment compensation based on sequential estimation," *IEEE Signal Processing Letters*, vol. 5, no. 3, March 1998.

[3] N. S. Kim, "Imm–based estimation for slowly evolving environments," *IEEE Signal Processing Letters*, vol. 5, no. 6, pp. 146–149, June 1998.

[4] K. Yao, B. E. Shi, P. Fung, and Z. Cao, "Residual noise compensation for robust speech recognition in nonstationary noise," in *ICASSP*, 2000, vol. 2, pp. 1125–1128.

[5] K. Yao, K.K. Paliwal, B. E. Shi, and S. Nakamura, "Noise compensation by a sequential kullback proximal algorithm," in *Inter. Workshop on Hands-Free Speech Recognition*, ATR, Kyoto, Japan, 2001.

[6] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden markov model parameters based on the kullback-leibler information measure," *IEEE Trans. on Signal Processing*, vol. 41, no. 8, August 1993.

[7] K. Yao, B. E. Shi, K.K.Palwal, and S. Nakamura, *A Sequential Kullback Proximal Algorithm for Parameter Estimation with Application to Sequential Noise Compensation*, ATR Spoken Language Translation Research Labs., 2001.

[8] S. Chréten and A. O. Hero III, "Kullback proximal point algorithms for maximum–likelihood estimation," *IEEE. Trans. on IT*, vol. 46, no. 5, August 2000.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Royal Stat. Soc.*, vol. 3g, pp. 1–38, 1977.

[10] D. Pearce, "Aurora project: Experimental framework for the performance evaluation of distributed speech recognition front-ends," in *ISCA ITRW ASR2000*, Sep. 2000.