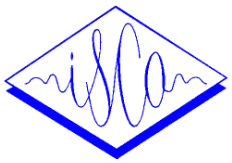

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

RESIDUAL NOISE COMPENSATION BY A SEQUENTIAL EM ALGORITHM FOR ROBUST SPEECH RECOGNITION IN NONSTATIONARY NOISE

Kaisheng YAO[⋆], *Bertram E. SHI*[†], *Satoshi NAKAMURA*[⋆], *and Zhigang CAO*[‡]

[⋆]ATR Spoken Language Translation Research Laboratories
2-2-2, Hikaridai Seika-cho, Souraku-gun, Kyoto, Japan 619-0288
[†]Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[‡]Department of Electronic Engineering, Tsinghua University, Beijing, P.R. China

ABSTRACT

We model noise as a stationary component plus a time varying residual. The stationary part is estimated off-line and compensated using Log-Add noise compensation. The time varying residual is estimated and compensated using a sequential EM algorithm. The residual noise compensation proceeds in parallel with the recognition process. Experimental results demonstrate that the proposed algorithm improves the recognition performance not only in highly nonstationary noise but also in slow-varying noise, compared with Log-Add noise compensation alone.

1. INTRODUCTION

Most approaches for model-based compensation for robust speech recognition in additive noise environments assume that the noise is stationary. Recently, researchers have begun to examine algorithms for the compensation of nonstationary noise [3] [4].

As a further work by the authors [4], we propose a sequential EM algorithm for residual noise compensation. Instead of a set of Kalman filters which was used in our previous work [4], we estimate the residual noise at each state by the sequential EM algorithm. The estimated residual noise is used to calculate the residual likelihood.

As stated in [4], we assume that noise effects can be separated into two parts. One part represents stationary noise effects, which can be compensated using a variant of parallel model combination (PMC) [2] prior to the recognition process. The other part represents effects from the residual time varying components of the noise. The estimation and compensation of the residual noise can be done in parallel with the recognition process.

The algorithm was evaluated using noise-corrupted utterances of TI connected digits. Our results indicated that the residual noise compensation could improve the recognition accuracy over that achieved by Log-Add noise compensation alone in not only highly nonstationary noise but also slow-varying noise.

2. RESIDUAL NOISE COMPENSATION

2.1. Observation Function

We apply residual noise compensation to models trained on Mel-scaled Frequency Cepstral Coefficients (MFCCs). We denote the linear-spectral power at filter bank j for clean speech and contaminating additive noise as $\sigma^2_{sj}(t)$ and $\sigma^2_{nj}(t)$, respectively. Assuming independence between the speech and additive noise, in the linear-spectral domain, the observation function for the linear-spectral power of the contaminated speech $O_j(t)$ is given as,

$$O_j(t) = \sigma^2_{sj}(t) + \sigma^2_{nj}(t) \tag{1}$$

In the log-spectral domain, we have [4],

$$O^l_j(t) = \mu^l_{sj}(t) + \log(1 + \exp(\mu^l_{nj}(t) - \mu^l_{sj}(t))) \tag{2}$$

where t denotes the time index and the superscript l represents the log-spectral domain. $\mu^l_{qj}(t) = \log(\sigma^2_{qj}(t))$, q is either s or n.

We express the noise as a combination of a part that is constant through the utterance and a $residual$ in the log-spectral domain. That is,

$$\mu^l_{nj}(t) = \mu^l_{nj} + \Delta\mu^l_{nj}(t) \tag{3}$$

As a result, at mixture m in state i, the observation function linearized around μ^l_{nj} is given as [4],

$$\begin{aligned} O^l_{imj}(t) &= \tilde{\mu}^l_{imj}(t) + C_{imj}(t)\Delta\mu^l_{nj}(t) \\ &+ O(\Delta\mu^l_{nj}(t)) \end{aligned} \tag{4}$$

where $C_{imj}(t) = \frac{\exp(\mu^l_{nj} - \mu^l_{imj}(t))}{1+\exp(\mu^l_{nj} - \mu^l_{imj}(t))}$. $\tilde{\mu}^l_{imj}(t)$ which represents the stationary noisy observation at mixture m in state i, is obtained as the sum of clean speech observation $\mu^l_{imj}(t)$ and a stationary noise contamination term, $\log(1+\exp(\mu^l_{nj} - \mu^l_{imj}(t)))$. $O(\Delta\mu^l_{nj}(t))$ contains higher-order terms of residual noise, which we assume are neglected in the following.

2.2. Model Representation of Observation Function

Under the assumptions that 1) there is no correlation between residual noise $\Delta\mu^l_{nj}(t)$ and stationary noise parameter μ^l_{nj} and that 2) noise $\mu^l_{nj}(t)$ and clean speech $\mu^l_{imj}(t)$ are independent, we can estimate the first- and second-order short-term statistics of cepstral features by the following equations,

$$\begin{aligned}\hat{\mu}_{\mathbf{im}}(\mathbf{t}) &= \hat{\mu}_{\mathbf{im}} + \mathbf{DCT^T}[\mathbf{E}[\mathbf{C_{im}}(\mathbf{t})^\mathbf{T}\mathbf{\Delta}\mu^\mathbf{l}_\mathbf{n}(\mathbf{t})]] \quad (5)\\ \hat{\mathbf{\Sigma}}_{\mathbf{im}}(\mathbf{t}) &= \hat{\mathbf{\Sigma}}_{\mathbf{im}} \\ &+ \mathbf{DCT^T}[\mathbf{E}[\mathbf{C^T_{im}}(\mathbf{t})\mathbf{\Delta}\mu^\mathbf{l}_\mathbf{n}(\mathbf{t})^\mathbf{T} \\ &\quad \mathbf{\Delta}\mu^\mathbf{l}_\mathbf{n}(\mathbf{t})\mathbf{C_{im}}(\mathbf{t})]]\mathbf{DCT} \quad (6)\end{aligned}$$

where the superscript T represents the transpose. $\mathbf{DCT}$ is a Discrete Cosine Transform (DCT) matrix. $E[\cdot]$ is the expectation over frames. $\hat{\mu}_{\mathbf{im}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{im}}$ are constants to the utterance, which are the mean and variance in the cepstral domain, respectively, compensated by algorithms such as PMC [2] prior to the recognition process. $\mathbf{\Delta}\mu^\mathbf{l}_\mathbf{n}(\mathbf{t})$ is a vector with the element of $\Delta\mu^l_{nj}(t)$.

To make an HMM based recognizer cope with time-varying noise, the compensated mean and variance of the model should incorporate time-varying parameters. Accordingly, $\mathbf{\Delta}\mu^\mathbf{l}_\mathbf{n}(\mathbf{t})$ is kept in the following in order to have a time-varying noise compensated model.

Let's further assume that 1) we have an estimation of log-spectral power $\{\mu^l_{nj}, j = 1 \cdots J\}$ for the stationary part of the noise, where J is the total number of filter banks, 2) there is no correlation between residual noise $\Delta\mu^l_{nj}(t)$ and stationary noisy observation $\tilde{\mu}^l_{imj}(t)$, and 3) the residual noise $\{\Delta\mu^l_{nj}(t), j = 1 \cdots J\}$ is independent from each other. From the above assumptions, the kth element in Equation 5 can be written as,

$$\hat{\mu}_{imk}(t) = \hat{\mu}_{imk} + \sum_j c_{kj} C_{imj} \Delta\mu^l_{nj}(t) \quad (7)$$

where c_{kj} is an inverse Discrete Cosine Transform (IDCT) coefficient.

If we additionally assume that the residual noise effects on the variance of the model can be neglected, the element in Equation 6 is given as,

$$\hat{\Sigma}_{imk}(t) = \hat{\Sigma}_{imk} \quad (8)$$

As a result, the likelihood score of the compensated model can be linearized only with respect to residual noise $\Delta\mu^l_{nj}(t)$.

2.3. Linearized Likelihood Score

The log-likelihood score at mixture m in hidden state i for cepstral observation $\mathbf{O^c}(\mathbf{t})$ is given by,

$$\begin{aligned}\hat{b}_{im}(\mathbf{O^c}(\mathbf{t})) &= \log(p(\mathbf{O^c}(\mathbf{t})|\mathbf{\Delta}\mu^\mathbf{l}_\mathbf{nj}(\mathbf{t}), \mathbf{i}, \mathbf{m})) \\ &= -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{k=1}^{K}\log(\hat{\Sigma}^2_{imk}) \\ &- \frac{1}{2}\sum_{k=1}^{K}\frac{(O^c_k(t) - \hat{\mu}_{imk}(t))^2}{\hat{\Sigma}^2_{imk}} \quad (9)\end{aligned}$$

where K is the feature vector size of static MFCC.

By the linearization of the likelihood score around $\mathbf{\Delta}\mu^\mathbf{l}_\mathbf{n}(\mathbf{t}) = \mathbf{0}$, we have a linearized likelihood score, which is a combination of the stationary noise component and the time varying residual noise, given as follows,

$$\hat{b}_{im}(\mathbf{O}^c(t)) = \hat{b}_{im}(\mathbf{O}^c(t))|_{\Delta\mu^l_n(t)=0} + \zeta_{im}(t) \quad (10)$$

where $\zeta_{im}(t)$ is denoted as the $residual\ likelihood$ afterwards, which is given as,

$$\begin{aligned}\zeta_{im}(t) &= \sum_{j=1}^{J}\zeta_{imj}(t)\Delta\mu^l_{nj}(t) + \frac{1}{2}\sum_{j=1}^{J}\zeta'_{imj}(t)(\Delta\mu^l_{nj}(t))^2 \\ \zeta_{imj}(t) &= \sum_{k=1}^{K}\frac{O^c_k(t) - \hat{\mu}_{imk}}{\Sigma^2_{imk}}c_{kj}C_{imj} \quad (11)\\ \zeta'_{imj}(t) &= \sum_{k=1}^{K}\frac{-1}{\Sigma^2_{imk}}(c_{kj}C_{imj})^2 \quad (12)\end{aligned}$$

The linearization of likelihood scores can make the stationary part and residual part be compensated separately. By the linearization of the likelihood score, the residual noise compensation can be done efficiently by only compensating the residual likelihood during the recognition process. This suggestion assumes that we can ignore the effects of the residual noise on the variance of the model.

3. SEQUENTIAL EM ALGORITHM

The EM algorithm [1] is an iterative method for finding ML parameter estimates. It works with complete data and iterates between estimating the log-likelihood of the complete data using incomplete data and the current parameter estimate (E-step) and maximizing the estimated log-likelihood function to obtain the new parameter estimate (M-step).

Specifically, at each mixture m in hidden state i, let observation $\mathbf{O^c}(\tau)$, previously estimated residual noise $\Lambda_{\tau-1} = \{\Delta\hat{\mu}^l_{nj}(t)|1 \le t \le \tau-1\}$ and currently estimated residual noise $\Delta\mu^l_{nj}(\tau)$ be the complete data. At time τ, for incomplete data $\mathbf{O^c}(\tau)$ and $\Lambda_{\tau-1}$, the auxiliary function for estimating residual noise $\Delta\mu^l_{nj}(\tau)$ at state i is given as,

$$Q_i(\Lambda_\tau, \Delta\mu^l_{nj}(\tau)) = \sum_{t=1}^{\tau}\varepsilon^{\tau-t}L_{it}(\mathbf{O^c}(\mathbf{t}), \mathbf{\Delta}\mu^\mathbf{l}_\mathbf{nj}(\mathbf{t})) \quad (13)$$

where ε is a forgetting factor with a range of 0.0 to 1.0.

$$L_{it}(\mathbf{O^c}(\mathbf{t}), \mathbf{\Delta}\mu^\mathbf{l}_\mathbf{nj}(\mathbf{t})) = \sum_{m=1}^{M}p(i, m|\mathbf{O^c}(\mathbf{t}), \mathbf{\Delta}\hat{\mu}^\mathbf{l}_\mathbf{nj}(\mathbf{t-1}))$$

$$[\log(p(\mathbf{O^c(t)}|\mathbf{\Delta\mu^l_{nj}(t)}, \mathbf{i}, \mathbf{m}) + \log(p(i,m))] \quad (14)$$

where M is the total mixture number in state i.

According to Equation 9 and 10, we can expand the likelihood in the brackets of Equation 14 around the previously estimated residual noise, $\Delta\hat{\mu}^l_{nj}(t-1)$. As a result, at state i, we get the residual likelihood as,

$$\begin{aligned}\zeta_{im}(t) &= \sum_{j=1}^{J} \zeta_{imj}(t)\Delta\hat{\mu}^l_{nj}(t-1) \\ &+ \frac{1}{2}\sum_{j=1}^{J} \zeta'_{imj}(t)(\Delta\hat{\mu}^l_{nj}(t-1))^2 \\ &+ \sum_{j=1}^{J} \eta_{imj}(t)(\Delta\mu^l_{nj}(t) - \Delta\hat{\mu}^l_{nj}(t-1)) \\ &+ \frac{1}{2}\sum_{j=1}^{J} \zeta'_{imj}(t)(\Delta\mu^l_{nj}(t) - \Delta\hat{\mu}^l_{nj}(t-1))^2\end{aligned} \quad (15)$$

where

$$\eta_{imj}(t) = \zeta_{imj}(t) + \zeta'_{imj}(t)\Delta\hat{\mu}^l_{nj}(t-1) \quad (16)$$

By maximizing the auxiliary function with respect to $\Delta\mu^l_{nj}(\tau)$ sequentially, at state i, we have the M-step given as,

$$\Delta\hat{\mu}^l_{nj}(\tau) = \Delta\hat{\mu}^l_{nj}(\tau-1) + \frac{S_{ij}(\tau)}{\Gamma_{ij}(\tau)} \quad (17)$$

where

$$S_{ij}(\tau) = \frac{\partial Q_i(\Lambda_\tau, \Delta\mu^l_{nj}(\tau))}{\partial\Delta\mu^l_{nj}(\tau)} \quad (18)$$

$$\Gamma_{ij}(\tau) = -\frac{\partial^2 Q_i(\Lambda_\tau, \Delta\mu^l_{nj}(\tau))}{\partial\Delta\mu^l_{nj}(\tau)^2} \quad (19)$$

They can be estimated by the following equations,

$$S_{ij}(\tau) = \sum_{m=1}^{M} p(i,m|\mathbf{O^c}(\tau), \mathbf{\Delta\hat{\mu}^l_n}(\tau-\mathbf{1}))\eta_{\mathbf{imj}}(\tau) \quad (20)$$

$$\begin{aligned}\Gamma_{ij}(\tau) &= \varepsilon\cdot\Gamma_{ij}(\tau-1) \\ &- \sum_{m=1}^{M} p(i,m|\mathbf{O^c}(\tau), \mathbf{\Delta\hat{\mu}^l_n}(\tau-\mathbf{1}))\zeta'_{\mathbf{imj}}(\tau)\end{aligned} \quad (21)$$

The sequential EM algorithm works independently at each state to estimate residual noise, $\Delta\hat{\mu}^l_{nj}(\tau)$, by iterative calculation of Equation 17 to Equation 21. The likelihood due to the residual noise is estimated by substituting $\Delta\mu^l_{nj}(\tau)$ with the estimated residual noise in Equation 11.

4. EXPERIMENTS ON THE SEQUENTIAL EM ALGORITHM

4.1. Experiment setup

Speaker-Independent TI-Digits recognition experiments were carried out with a Viterbi recognizer to test the noise compensation approach. The digits models and background noise model were trained on clean speech utterances using the HMM toolkit (HTK). The contaminated speech for the test was generated by artificially adding different levels of noise to the clean speech. All noise signals were from a Noisex-92 database.

Five hundred connected digits utterances from 15 speakers and 100 connected digits utterances from four speakers unseen in the training set were used for training and testing, respectively. There were 11 whole word models for 10 digits (zero is pronounced as oh or zero) and one silence model. Each digit was modeled by a four-Gaussian-mixture 10-state (including a nonemitting initial and final state) left-to-right HMM without skip states. Gaussian output probability distributions with diagonal covariance matrices were used for each state. The silence model was a four-Gaussian-mixture 3-state (with a first and last nonemitting state) HMM.

The speech signals were down-sampled from 20kHz to 16kHz. The window size was 25.0ms with a 10.0ms shift. Twenty six filters were used in the binning stage. The features were the static MFCC with the dynamic MFCC.

4.2. Results

Only the mean of the static MFCC was compensated. The stationary noise statistic was calculated from five seconds of contaminating noise before the recognition process. Noise statistic was modeled by a one-mixture Gaussian state. Stationary noise compensation was carried out using Log-Add noise compensation.

The following tables show the performance in White, Babble and Machinegun noise. These noise situation are representative of stationary, slow-varying and highly nonstationary noise, respectively. Since the sequential algorithm is iterative in nature, the initial condition might be important to the performance. In this paper, we compare the performance with three different sets of $\Gamma(0)$, where it is set the same for all of the log-spectral indices in each condition.

As can be seen from the tables, the Log-Add noise compensation contributes much towards the system robustness to noise. The averaged Word Error Rate Reductions are about 77.8%, 65.4%, 32.2% for the White, Babble, and Machinegun noise, respectively.

We show the performance of the residual noise compensation with the best forgetting factor in the tables. As further evidenced from the tables, in the Babble and Machinegun noise, there are further word error rate reductions

by using the residual noise compensation, compared with the reductions achieved by the Log-Add noise compensation. For example, in 33.1 dB Machinegun noise, the word error rate drops from 13.3% to 8.7%. Although the word error rate reduction is consistent, we can observe that there are no error rate reductions in some lower SNR ranges in the noises. We believe that, by using more complex noise models, the residual noise compensation can achieve further error rate reductions in lower SNR ranges. We can also see that the initial set of $\Gamma(0)$ does not have effects on the performance when we set the forgetting factor equal to 0.98.

As a whole, the residual noise compensation by the sequential EM algorithm with a forgetting factor of 0.98 can have a consistent performance improvement over a system compensated by the Log-Add noise compensation, not only in slow-varying noise, but also in highly non-stationary noise.

Table 1: Word Error Rate (in %) in White noise environments. Baseline denotes the performance without noise compensation. LAdd denotes Log-Add noise compensation. SEM(120.0), SEM(2.0) and SEM(1.0) each denote the performance of the sequential EM algorithm with a different initial set of $\Gamma(0)$ equal to 120.0, 2.0, and 1.0, respectively. $\varepsilon = 0.98$.

SNR (dB)	8.8	16.0	20.4	40.4
Baseline	80.0	71.3	62.3	30.0
LAdd	30.0	13.7	9.0	5.3
SEM (1.0)	30.0	13.7	9.0	5.3
SEM (2.0)	30.0	13.7	9.0	5.3
SEM (120.0)	30.0	13.7	9.0	5.3

Table 2: Word Error Rate (in %) in Babble noise environments. Baseline denotes the performance without noise compensation. LAdd denotes Log-Add noise compensation. SEM(120.0), SEM(2.0) and SEM(1.0) each denote the performance of the sequential EM algorithm with a different initial set of $\Gamma(0)$ equal to 120.0, 2.0, and 1.0, respectively. $\varepsilon = 0.98$.

SNR (dB)	0.7	9.4	12.9	32.6
Baseline	86.3	90.0	84.0	32.3
LAdd	55.0	31.7	19.3	5.3
SEM (1.0)	55.0	31.7	19.3	5.0
SEM (2.0)	55.0	31.7	19.3	5.0
SEM (120.0)	55.0	31.7	19.3	5.0

Table 3: Word Error Rate (in %) in Machinegun noise environments. Baseline denotes the performance without noise compensation. LAdd denotes Log-Add noise compensation. SEM(120.0), SEM(2.0) and SEM(1.0) each denote the performance of the sequential EM algorithm with a different initial set of $\Gamma(0)$ equal to 120.0, 2.0, and 1.0, respectively. $\varepsilon = 0.98$.

SNR (dB)	1.4	8.6	13.1	33.1
Baseline	59.0	49.3	47.0	16.3
LAdd	33.3	32.0	32.0	13.3
SEM (1.0)	33.3	30.7	30.0	8.7
SEM (2.0)	33.3	30.7	30.0	8.7
SEM (120.0)	33.3	30.7	30.0	8.7

5. CONCLUSION

We have presented a sequential EM algorithm for residual noise compensation. The sequential EM algorithm estimates residual noise parameters, and their effects are compensated by residual noise compensation. Through a set of experiments, we showed that the residual noise compensation algorithm with a proper forgetting factor can have consistent word error rate reductions over a system compensated by Log-Add noise compensation alone, not only in slow-varying noise, but also in highly nonstationary noise.

6. REFERENCES

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

[2] M. J. F. Gales and S. J. Young. Robust speech recognition in additive and convolutional noise using parellel model combination. *Computer, Speech and Language*, 9:289–307, 1995.

[3] Nam Soo Kim. Imm–based estimation for slowly evolving environments. *IEEE Signal Processing Letters*, 5(6):146–149, June 1998.

[4] Kaisheng Yao, Bertram E. Shi, Pascale Fung, and Zhigang Cao. Residual noise compensation for robust speech recognition in nonstationary noise. In *Proceeding of ICASSP*, volume 2, pages 1125 – 1128, 2000.