Model-based approach for robust speech recognition in noisy environments with multiple noise sources *

Do Yeong Kim[1,2], Nam Soo Kim[2], Chong Kwan Un[1]

[1]Department of Elec. Eng., KAIST, Korea
dykim@eekaist.kaist.ac.kr
[2]Human and Computer Interaction Lab., SAIT, Korea
nskim@green.sait.samsung.co.kr

ABSTRACT

In this paper, we consider the hidden Markov model(HMM) parameter compensation in noisy environments with multiple noise sources based on the vector Taylor series(VTS) approach. General formulations for multiple environmental variables are derived and systematic *expectation-maximization*(EM) solutions are presented in maximum likelihood(ML) sense. It is assumed that each noise source is independent and having Gaussian distribution. To evaluate proposed method, we conduct speaker independent isolated word recognition experiments in various noisy environments. Experimental results show that proposed algorithm ahieves significant improvement. Especially, the proposed method is consistently more effective than the parallel model combination(PMC) based on log-normal approximation.

1 Introduction

Presently, problems with noise-robustness is one of most important issues in speech recognition. Various methods have been proposed, such as robust distance measures, feature vector transformation and model parameter adaptation. Feature vector transformation using signal Gaussian mixture achieve successful results in the log-spectral and cepstral domains[1][2][3]. Also, model parameter adaptation algorithms affected by speaker adaptation schemes show more improved performance over feature vector transformation. In time-varying noisy condition, however, fast adaptation is required and use of sufficient adaptation speech for adjusting all model parameters is difficult, while most speaker adaptation schemes use some quantity of adaptation data.

Moreno proposed the vector Taylor series(VTS) approach to formulate relation between clean and noisy speech signals and analytically solve the noise-robust speech processing in the feature vector transform domain[2]. He achieved significant improvement compared with other methods. Using only 2 environmental variables, i.e., additive noise and spectral tilt, the VTS approach yields reliable performance. But, it was performed in the log-spectral and assumed independence between log-spectral elements for reduction of computational burden. Since many speech recognition systems accept cepstral coefficients as a feature vector, compensation in the log-spectral domain requires additional condition such as log-normal approximation. To solve these problems, we generalized the VTS algorithm and applied it to cepstral domain in our previous work. We presented an exact *expectation-maximization*(EM) solution of VTS with noise statistics[4]. Also, we developed new model parameter adaptation algorithm based on the VTS[5].

In this paper, we consider speech recognition in noisy environments with multiple noise sources. General formulations for multiple environmental variables are derived and systematic EM solutions are presented in the maximum likelihood(ML) sense. It is assumed that each noise source is independent and having Gaussian distribution.

2 Environment modeling

2.1 Modeling additive noise

Let us consider simple additive noise environment. Corrupted speech signal (or feature vector) $\mathbf{y}$ can be expressed as:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \mathbf{n}) \tag{1}$$

where $\mathbf{x}$ is clean speech signal, and $\mathbf{n}$ is parameter that represents the effects of the additive noise. In general, the generic function $\mathbf{f}(\mathbf{x}, \mathbf{n})$ is nonlinear and defined by the parameter domain. For example, if we assume that all parameters are defined in the logarithmic domain, we get

$$\mathbf{y}_l = \log(\exp(\mathbf{x}_l) + \exp(\mathbf{n}_l)) \tag{2}$$

where l denotes the log-spectral parameter.

*This work was partially supported by Samsung Advanced Institute of Technology(SAIT).

In the cepstral domain,

$$\mathbf{y}_c = C\log(\exp(C^{-1}\mathbf{x}_c) + \exp(C^{-1}\mathbf{n}_c)) \quad (3)$$

or

$$\mathbf{y}_c = \mathbf{x}_c + C\log(\exp(C^{-1}(\mathbf{n}_c - \mathbf{x}_c))) \quad (4)$$

as Acero used in [1], where c representing cepstral parameters, C denoting discrete cosine transform(DCT) matrix, and C^{-1} being inverse DCT matrix. Also, the right-hand side of eq. (2)-(4) represent contamination procedure defined by parameter domain.

There is no closed form solution for mean and variance of corrupted speech signal, $\mathbf{y}$, in eq. (2)-(4). To get exact solution, numerical integration was performed in several previous studies, but, it isn't practical because of its heavy computational burden.

2.2 Truncated VTS approximation

Moreno proposed the VTS approach by which nonlinear comtamination function was approximated as truncated vector Talyor series [2]. Let $\mathbf{y} = [y_1, y_2, \cdots, y_N]'$ be a noisy cepstral feature vector [1] with dimension N. Assume that $\mathbf{y}$ is related to the clean feature $\mathbf{x} = [x_1, x_2, \cdots, x_N]'$, and additive noise $\mathbf{n} = [n_1, n_2, \cdots, n_N]'$ by

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \underbrace{\begin{bmatrix} f_1(\mathbf{x},\mathbf{n}) \\ f_2(\mathbf{x},\mathbf{n}) \\ \vdots \\ f_N(\mathbf{x},\mathbf{n}) \end{bmatrix}}_{\mathbf{f}(\mathbf{x},\mathbf{n})} \quad (5)$$

in which f_1, f_2, . . ., f_N represent the contamination procedure under consideration. By expanding VTS around $(\mathbf{x}_0, \mathbf{n}_0)$ and taking only upto the first-order terms, we can approximate (5) such that

$$\mathbf{y} \simeq (\nabla_{\mathbf{x}}\mathbf{f})'\mathbf{x} + (\nabla_{\mathbf{n}}\mathbf{f})'\mathbf{n} + \mathbf{g}(\mathbf{x}_0, \mathbf{n}_0) \quad (6)$$

where

$$(\nabla_{\mathbf{x}}\mathbf{f})' = \begin{bmatrix} (\nabla_{\mathbf{x}} f_1)' \\ (\nabla_{\mathbf{x}} f_2)' \\ \vdots \\ (\nabla_{\mathbf{x}} f_N)' \end{bmatrix}, \quad (\nabla_{\mathbf{n}}\mathbf{f})' = \begin{bmatrix} (\nabla_{\mathbf{n}} f_1)' \\ (\nabla_{\mathbf{n}} f_2)' \\ \vdots \\ (\nabla_{\mathbf{n}} f_N)' \end{bmatrix}$$

and

$$\mathbf{g}(\mathbf{x}_0, \mathbf{n}_0) = \begin{bmatrix} g_1(\mathbf{x}_0, \mathbf{n}_0) \\ g_2(\mathbf{x}_0, \mathbf{n}_0) \\ \vdots \\ g_N(\mathbf{x}_0, \mathbf{n}_0) \end{bmatrix}. \quad (7)$$

In [4], detail procedure for environmental variable estimation can be found when there exist noise statistics, and also we developed a method to estimate not only additive noise but also spectral tilt and additive noise variance using the EM algorithm[5].

[1] for brevity, we drop the subscript c.

2.3 Noisy environment of multiple noise sources

Even though there are multiple noise sources, and each source has its statistics, we can apply VTS approach. It is assumed there are M independent noise sources, and each noise source is a Guassian. Also, we assume that we know contamination procedure exactly. (Contamination procedure is generally nonlinear and may be extremely complex.)

Using truncated VTS, we can get following approximation

$$\mathbf{y} \simeq (\nabla_{\mathbf{x}}\mathbf{f})'\mathbf{x} + \sum_{m=1}^{M}(\nabla_{\mathbf{n}^m}\mathbf{f})'\mathbf{n}^m + \mathbf{g}(\mathbf{x}_0, \mathbf{n}_0^1, \cdots, \mathbf{n}_0^M) \quad (8)$$

where $\mathbf{n}^m$ denotes a noisy feature vector from m th noise source.

We assume that the probability density function(PDF) of speech signal can be represented by a summation of multivariate Gaussian distributions :

$$p(\mathbf{x}) \sim \sum_{k=1}^{K} p(k)N(\mathbf{x} : \mu_k, \Sigma_k) \quad (9)$$

where K is the total number of mixture components and $p(k), \mu_k, \Sigma_k$ represent given a priori probability, mean and variance of k-th Gaussian distribution, respectively.

To obtain re-estimation fomulars, consider an auxiliary function given by

$$Q(\bar{\lambda}, \lambda) = E[\log p(\mathbf{X}, \mathbf{N}^1, \mathbf{N}^2, \cdots, \mathbf{N}^M, \mathbf{K}|\lambda)|\mathbf{Y}, \bar{\lambda}]$$

where $\mathbf{N}^m = \{\mathbf{n}_1^m, \mathbf{n}_2^m, \cdots, \mathbf{n}_T^m\}$, $1 \le i \le M$, denotes the m th noise vector sequence which is statistically independent of the clean feature vector sequence and other noise vector sequences, and M is the number of noise sources. T is a length of vector sequence, and $\mathbf{K} = \{k_1, k_2, \cdots, k_T\}$ is a hidden sequence of mixture components. Given $\bar{\lambda}$, new parameter estimates, $\hat{\lambda}$ are sought according to

$$\hat{\lambda} = \arg\max_{\lambda} Q(\bar{\lambda}, \lambda).$$

Assuming that each noise source is a Gaussian, we take the gradient of $Q(\bar{\lambda}, \lambda)$ with respect to $\mu_{\mathbf{n}}^m$, mean vector of m-th noise source. Equating the gradient to zero, we can get re-estimation equation of m th noise mean as follows

$$\hat{\mu}_{\mathbf{n}}^m = \frac{1}{T}\sum_t \sum_k p(k|\mathbf{y}_t, \bar{\lambda})E[\mathbf{n}_t^m|\mathbf{y}_t, k, \bar{\lambda}]. \quad (10)$$

In a similar manner, we can also re-estimate m-th noise variance,

$$\hat{\Sigma}_{\mathbf{n}}^m = \frac{1}{T}\sum_t \sum_k p(k|\mathbf{y}_t, \bar{\lambda})E[\mathbf{n}_t^m \mathbf{n}_t^{m\prime}|\mathbf{y}_t, k, \bar{\lambda}] - \hat{\mu}_{\mathbf{n}}^m \hat{\mu}_{\mathbf{n}}^{m'}. \quad (11)$$

More detail explanations are given in the Appendix.

2.4 HMM model parameter compensation

For model parameter compensation without adaptation speech, we need to find

$$(\hat{\lambda}, \hat{\mathbf{W}}) = \arg \max_{(\lambda, \mathbf{W})} p(\mathbf{Y}, \mathbf{W}|\lambda, \Lambda_{\mathbf{X}}) = \arg \max_{(\lambda, \mathbf{W})} p(\mathbf{Y}, \mathbf{W}|\lambda, \Lambda_{\mathbf{X}}) p(\mathbf{W}) \quad (12)$$

where $\mathbf{W} = \{W_1, W_2, \cdots, W_3\}$ is a word sequence embedded in $\mathbf{Y}$, $\Lambda_{\mathbf{X}}$ is a model parameter set of clean speech. $\mathbf{W}$ and λ are jointly maximized by keeping λ fixed and maximizing over $\mathbf{W}$, and the keeping $\mathbf{W}$ fixed and mximizing over λ iteratively.

After several steps similar to previous section, we get the following equations.

$$\hat{\mu}_{\mathbf{n}}^m = \frac{1}{T} \sum_t \sum_i \sum_j \gamma_t(i,j) E[\mathbf{n}_t^m | \mathbf{y}_t, i, j, \bar{\lambda}], \quad (13)$$

$$\hat{\Sigma}_{\mathbf{n}}^m = \frac{1}{T} \sum_t \sum_i \sum_j \gamma_t(i,j) E[\mathbf{n}_t^m \mathbf{n}_t^{m\prime} | \mathbf{y}_t, i, j, \bar{\lambda}] - \hat{\mu}_{\mathbf{n}}^m \hat{\mu}_{\mathbf{n}}^{m'}. \quad (14)$$

where $\gamma_t(i,j) = p(\mathbf{Y}, s_t = n, c_t = m|\bar{\lambda})$ is the joint likelihood of $\mathbf{Y}$ and the m-th mixture component of the n-th state with $\bar{\lambda}$ producing the observation $\mathbf{y}_t$.

By approximation of truncated VTS given eq. (8), we finally get following new hidden Markov model(HMM) parameters,

$$\mu_{\mathbf{y},i,j} = (\nabla_{\mathbf{x}} \mathbf{f})' \mu_{\mathbf{x},i,j} + \sum_{m=1}^{M} (\nabla_{\mathbf{n}^m} \mathbf{f})' \mu_{\mathbf{n}}^m + \mathbf{g}(\mu_{\mathbf{x},i,j}, \mu_{\mathbf{n}}^1, \cdots, \mu_{\mathbf{n}}^m) \quad (15)$$

$$\Sigma_{\mathbf{y},i,j} = (\nabla_{\mathbf{x}} \mathbf{f})' \Sigma_{\mathbf{x},i,j} (\nabla_{\mathbf{x}} \mathbf{f}) + \sum_{m=1}^{M} (\nabla_{\mathbf{n}^m} \mathbf{f})' \Sigma_{\mathbf{n}}^m (\nabla_{\mathbf{n}^m} \mathbf{f}) \quad (16)$$

where $\mu_{\mathbf{y},i,j}$ and $\Sigma_{\mathbf{y},i,j}$ are noisy mean and variance of i-th state, j-th mixture, and $\mu_{\mathbf{x},i,j}$, $\Sigma_{\mathbf{x},i,j}$ denote clean speech mean and variance of i-th state, j-th mixture.

3 Experiments

3.1 Task and database

Performances of the proposed methods were evaluated with speaker-independent isolated word recognition experiments. The vocabulary consists of 75 Korean phonetically-ballanced words. 90 male speakers uttered the words once to construct the database for training and evaluation. Utterances from 60 speakers constructed the training data and those from the other 30 speakers were used for evaluation. Each utterance was digitized with a sampling rate of 16kHz. A 18th-order mel-scaled log filterbank energy vector was extracted for every frame of 10 ms. By applying DCT, a 13th-order cepstral coefficient vector was derived for each frame and used for recognition. 32 phoneme models were used as the basic units of recognition. Each unit was modeled by a three-state continuous mixture HMM which is a simple left-to-right model without skipping where each state has three mixture components. 3 types of noise - Computer generated white Gaussian noise, NOISEX92 car noise(VOLVO), and NOISEX92 babble noise - were considered. According to various SNR, scaled noise samples were added to speech signal in time-domain.

3.2 Experimental results

We compensated HMM parameters according to changes of environments in these experiments. Any prior information was not used for on-line adaptation. To use as a reference, we implemented well-known parallel model combination(PMC) algorithm based on log-normal approximation[6]. Since noise samples were added to speech signal in time-domain, there was no explicit linear channel distortion in our experiments. But, variablities between speakers could be considered a kind of spectral tilt. Also, errors of assumed model could make other distortions. Thus, we assumed noisy environments with 2 noise sources, addtive noise and spectral tilt as other works[1][2]. 2 sources were modeled as independent Gaussian, respectively. Initial noise model parameters were obtained from short slience frames (3-4 frame) before beginning of speech.

When clean speech was applied, our system showed 93.4% recognition rate. Table 1. shows experimental result of speaker independent isolated word recognition in various noise environmetns. In all noisy condition, recognition performance of baseline system was degraded seriously when no compensation scheme is adopted. Especially additive white Gaussian(AWG) noise and BABBLE noise degraded performance drastically even at relatively high SNR(20dB). In all noisy condition of various SNR, our proposed method outperformed the well-known PMC algorithm. Note that it was effective to not only stationary noise (AWG, CAR) but also nonstationary noise (BABBLE).

4 Conclusions

In this paper, we presented a novel method to compensate HMM model parameters in noisy environments. Previous VTS algorithm was reviewed and extended to multiple noise source case. Environmental variables (mean and variance of noise sources) were estimated using the EM algorithm and detail procedure was presented for compensation of HMM parameters. Developed method did not use any prior information of noise source, and

Table 1: Experimental results of speaker independent isolated word recognition in various noise conditions(%).

Noise type	Comp. algo.	SNR (dB)			
		30	20	10	0
AWG	None	83.7	46.1	8.9	3.1
	PMC	91.1	85.2	71.3	38.3
	Proposed	92.1	87.5	77.0	49.8
CAR	None	93.3	92.7	88.5	66.3
	PMC	92.6	92.5	91.4	88.2
	Proposed	93.4	93.0	92.6	89.2
BABBLE	None	89.7	67.7	34.6	9.3
	PMC	89.3	82.2	62.7	27.5
	Proposed	92.3	87.3	73.4	40.9

only need utterance to be recognized. To evaluate proposed method, we performed speaker-independent isolated word recognition experiments. Proposed method outperformed well-known PMC algorithm at various condition. Especially, it effectively compensated the HMM parameters in the nonstationary BABBLE noise environment as well as stationary condition.

Appendix

Assuming that each noise source is a Gaussian, and taking the gradient of $Q(\bar{\lambda}, \lambda)$ with respect to $\mu^m_{\mathbf{n}}$, mean vector of m th noise source, we can obtain following formula.

$$\frac{\partial Q}{\partial \mu^m_{\mathbf{n}}} = \sum_t \sum_k \int_{\Psi_t} (\Sigma^m_{\mathbf{n}})^{-1}(\mathbf{n}^m_t - \mu^m_{\mathbf{n}}) p(\Psi_t, k|\mathbf{y}_t, \bar{\lambda}) d\Psi_t$$

where $\Psi_t = \{\mathbf{n}^1_t, \mathbf{n}^2_t, \cdots, \mathbf{n}^m_t\}$. Equating above equation to zero, and after several step we can get re-estimation formula as follows

$$\hat{\mu}^m_{\mathbf{n}} = \frac{1}{T}\sum_t \sum_k p(k|\mathbf{y}_t, \bar{\lambda}) \mu^m_{\mathbf{n}}(\mathbf{y}_t, k, \bar{\lambda}) \qquad (17)$$

in which

$$\begin{aligned} \mu^m_{\mathbf{n}}(\mathbf{y}_t, k, \bar{\lambda}) &= E[\mathbf{n}^m_t|\mathbf{y}_t, k, \bar{\lambda}] \\ &= \tilde{\Sigma}^m_{\mathbf{n}}(\bar{\Sigma}^m_{\mathbf{n}} + \tilde{\Sigma}^m_{\mathbf{n}})^{-1}\bar{\mu}^m_{\mathbf{n}} + \bar{\Sigma}^m_{\mathbf{n}}(\bar{\Sigma}^m_{\mathbf{n}} + \tilde{\Sigma}^m_{\mathbf{n}})^{-1}\tilde{\mu}^m_{\mathbf{n}} \end{aligned} \qquad (18)$$

where

$$\begin{aligned} \tilde{\mu}^m_{\mathbf{n}} &= (\nabla^m_{\mathbf{n}}\mathbf{f})^{-t}[\mathbf{y}_t - (\nabla_{\mathbf{x}}\mathbf{f})'\mu_k - \sum_{i\neq m}(\nabla^i_{\mathbf{n}}\mathbf{f})'\mu^i_{\mathbf{n}} \\ &\quad - \mathbf{g}(\mathbf{x}_0, \mathbf{n}^1_0, \cdots, \mathbf{n}^M_0)] \end{aligned} \qquad (19)$$

$$\tilde{\Sigma}^m_{\mathbf{n}} = (\nabla^m_{\mathbf{n}}\mathbf{f})^{-t}(\nabla_{\mathbf{x}}\mathbf{f})'\Sigma_k(\nabla_{\mathbf{x}}\mathbf{f})(\nabla^m_{\mathbf{n}}\mathbf{f})^{-1}. \qquad (20)$$

μ_k and Σ_k denote speech feature vector mean and variance of k th mixture (codeword), respectively. In a similar manner, we can get new variance of m th noise source.

$$\hat{\Sigma}^m_{\mathbf{n}} = \frac{1}{T}\Sigma_t \Sigma_k p(k|\mathbf{y}_t, \bar{\lambda}) E[\mathbf{n}^m_t \mathbf{n}^{m\prime}_t|\mathbf{y}_t, k, \bar{\lambda}] - \hat{\mu}^m_{\mathbf{n}}\hat{\mu}^{m'}_{\mathbf{n}}$$

where

$$\begin{aligned} & E[\mathbf{n}^m_t \mathbf{n}^{m\prime}_t|\mathbf{y}_t, k, \bar{\lambda}] \\ = \; & E[(\mathbf{n}_t - \mu_{\mathbf{n}}(\mathbf{y}_t, k, \bar{\lambda}))(\mathbf{n}_t - \mu_{\mathbf{n}}(\mathbf{y}_t, k, \bar{\lambda}))'|\mathbf{y}_t, k, \bar{\lambda}] \\ & + \mu_{\mathbf{n}}(\mathbf{y}_t, k, \bar{\lambda})\mu_{\mathbf{n}}(\mathbf{y}_t, k, \bar{\lambda})' \end{aligned} \qquad (21)$$

and

$$\begin{aligned} & E[(\mathbf{n}_t - \mu_{\mathbf{n}}(\mathbf{y}_t, k, \bar{\lambda}))(\mathbf{n}_t - \mu_{\mathbf{n}}(\mathbf{y}_t, k, \bar{\lambda}))'|\mathbf{y}_t, k, \bar{\lambda}] \\ = \; & [(\bar{\Sigma}^m_{\mathbf{n}})^{-1} + (\tilde{\Sigma}^m_{\mathbf{n}})^{-1}]^{-1}. \end{aligned} \qquad (22)$$

References

[1] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*, Kluwer academic publishers, 1993.

[2] P. J. Moreno, B. Raj and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," *Proc. of Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, GA, pp. 733-736, May 1996.

[3] B. Raj, E. B. Gouvea, P. J. Moreno and R. M. Stern, "Cepstral compensation by polynomial approximation for environment-independent speech recognition," *Proc. of Int. Conf. Spoken Language Processing*, Philadelphia, PA, pp. 2340-2343, Oct. 1996.

[4] N. S. Kim, D. Y. Kim, B. G. Kong and S. R. Kim, "Application of VTS to environment compensation with noise statistics," *Proc. of ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, Apr. 1997.

[5] D. Y. Kim, C. K. Un and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication*, submitted for publication.

[6] M. J. F. Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. Thesis, Univ. of Cambridge, 1995.