Robust Speech Recognition for
Unknown Communication Channels
Pont-à-Mousson, France
,April 17-18, 1997

APPLICATION OF VTS TO ENVIRONMENT COMPENSATION WITH NOISE STATISTICS

Nam Soo Kim *Do Yeong Kim* *Byung Goo Kong* *Sang Ryong Kim*

Human & Computer Interaction Lab.
Information Processing Sector
Samsung Advanced Institute of Technology
P. O. Box 111, Suwon 440-600, Korea
nskim@green.sait.samsung.co.kr

ABSTRACT

Recently, the vector Taylor series (VTS) approach was proposed as an efficient method for robust speech recognition under various environmental conditions. The VTS approach makes an approximation to the speech contamination procedure by a linearized model and estimates the parameter values using the *expectation-maximization* (EM) algorithm. In this paper, we apply the VTS approach to environment compensation with assumed noise statistics. In addition, we present a Bayesian adaptation technique with which we can incorporate the a priori knowledge about the noise statistics to the parameter estimation procedure.

1. INTRODUCTION

Environment compensation techniques which compensate the effects of added noise and channel distortion have been developed for robust speech recognition [1] - [3]. In general, these techniques make assumptions on both the clean speech distribution and the environmental contamination procedure, and by estimating the environmental parameters, transform the noisy speech features to the clean features. Among them, the vector Taylor series (VTS) approach is considered to be effective since it can approximate the highly nonlinear contamination procedure by a simplified linear model, which improves the mathematical tractability in parameter estimation [1], [2].

In the VTS approach, the clean speech features are characterized by a mixture of Gaussian distributions, and the contamination procedure is separately approximated for each mixture component by using the truncated VTS expansion. Environmental parameters such as the added noise feature and spectral tilt vectors are derived through the *expectation-maximization* (EM) algorithm which iteratively updates the parameter estimates.

In this paper, we present an environment compensation algorithm which is based on the VTS approach with noise statistics assumption. A Bayesian adaptation technique is also applied in order to incorporate the a priori knowledge of the noise statistics to the parameter estimation procedure. Furthermore, the performances of the presented approaches are evaluated through the speaker-independent isolated word recognition experiments.

2. TRUNCATED VTS APPROXIMATION

Let $\mathbf{z} = [z_1, z_2, \cdots, z_N]'$ be a noisy feature vector with dimension N. Assume that $\mathbf{z}$ is related to the clean feature $\mathbf{x} = [x_1, x_2, \cdots, x_N]'$, noise $\mathbf{n} = [n_1, n_2, \cdots, n_N]'$ and spectral tilt $\mathbf{q} = [q_1, q_2, \cdots, q_N]'$ by

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix} = \underbrace{\begin{bmatrix} f_1(\mathbf{n},\mathbf{x},\mathbf{q}) \\ f_2(\mathbf{n},\mathbf{x},\mathbf{q}) \\ \vdots \\ f_N(\mathbf{n},\mathbf{x},\mathbf{q}) \end{bmatrix}}_{\mathbf{f}(\mathbf{n},\mathbf{x},\mathbf{q})} \quad (1)$$

in which f_1, f_2, ..., f_N represent the contamination procedure. By using VTS expansion around $(\mathbf{n}_0, \mathbf{x}_0, \mathbf{q}_0)$ and taking only upto the first order terms, we can approximate (1) such that

$$\mathbf{z} = (\nabla_{\mathbf{x}}\mathbf{f})'\mathbf{x} + (\nabla_{\mathbf{n}}\mathbf{f})'\mathbf{n} + (\nabla_{\mathbf{q}}\mathbf{f})'\mathbf{q} + \mathbf{g}(\mathbf{n}_0, \mathbf{x}_0, \mathbf{q}_0) \quad (2)$$

where

$$(\nabla_{\mathbf{x}}\mathbf{f}) = \begin{bmatrix} \frac{\partial f_1(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial x_1} & \cdots & \frac{\partial f_N(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial x_N} & \cdots & \frac{\partial f_N(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial x_N} \end{bmatrix} \quad (3)$$

$$(\nabla_{\mathbf{n}}\mathbf{f}) = \begin{bmatrix} \frac{\partial f_1(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial n_1} & \cdots & \frac{\partial f_N(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial n_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial n_N} & \cdots & \frac{\partial f_N(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial n_N} \end{bmatrix} \quad (4)$$

$$(\nabla_{\mathbf{q}}\mathbf{f}) = \begin{bmatrix} \frac{\partial f_1(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial q_1} & \cdots & \frac{\partial f_N(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial q_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial q_N} & \cdots & \frac{\partial f_N(\mathbf{n}_0,\mathbf{x}_0,\mathbf{q}_0)}{\partial q_N} \end{bmatrix} \quad (5)$$

and

$$\mathbf{g}(\mathbf{n}_0, \mathbf{x}_0, \mathbf{q}_0) = \\ \mathbf{f}(\mathbf{n}_0, \mathbf{x}_0, \mathbf{q}_0) - (\nabla_{\mathbf{x}}\mathbf{f})'\mathbf{x}_0 - (\nabla_{\mathbf{n}}\mathbf{f})'\mathbf{n}_0 - (\nabla_{\mathbf{q}}\mathbf{f})'\mathbf{q}_0 \ . \tag{6}$$

We assume that $(\nabla_{\mathbf{x}}\mathbf{f})$, $(\nabla_{\mathbf{n}}\mathbf{f})$ and $(\nabla_{\mathbf{q}}\mathbf{f})$ satisfy the regularity condition which indicates that $(\nabla_{\mathbf{x}}\mathbf{f})^{-1}$, $(\nabla_{\mathbf{n}}\mathbf{f})^{-1}$ and $(\nabla_{\mathbf{q}}\mathbf{f})^{-1}$ exist for all possible $(\mathbf{n}_0, \mathbf{x}_0, \mathbf{q}_0)$ points.

For an example, let us consider the environment where the speech waveform is corrupted by an additive stationary noise and a linear channel. In the log-spectral domain this environment can be expressed by the relation

$$z_i = x_i + q_i + \log\left(1 + e^{n_i - x_i - q_i}\right) \quad \text{for } i = 1, 2, \cdots, N \tag{7}$$

where $\mathbf{z} = [z_1, z_2, \cdots, z_N]'$ represents the log power spectrum of the degraded speech, $\mathbf{x} = [x_1, x_2, \cdots, x_N]'$ is the log power spectrum of the clean speech, $\mathbf{n} = [n_1, n_2, \cdots, n_N]'$ is the log power spectrum of the additive noise and $\mathbf{q} = [q_1, q_2, \cdots, q_N]'$ denotes the log channel transfer function. Taking gradients, we can get

$$(\nabla_{\mathbf{x}}\mathbf{f}) = \mathrm{diag}\,[\xi_1, \xi_2, \cdots, \xi_N] \tag{8}$$

$$(\nabla_{\mathbf{q}}\mathbf{f}) = \mathrm{diag}\,[\xi_1, \xi_2, \cdots, \xi_N] \tag{9}$$

$$(\nabla_{\mathbf{n}}\mathbf{f}) = \mathrm{diag}\,[1 - \xi_1, 1 - \xi_2, \cdots, 1 - \xi_N] \tag{10}$$

where

$$\xi_i = \frac{1}{1 + e^{n_{0,i} - x_{0,i} - q_{0,i}}} \ . \tag{11}$$

Therefore, the relation in (7) is approximated by

$$z_i = \\ \xi_i(x_i - x_{0,i}) + \xi_i(q_i - q_{0,i}) + (1 - \xi_i)(n_i - n_{0,i}) \quad \text{for } i = 1, 2, \cdots, N \ . \tag{12}$$

3. VTS ALGORITHM WITH NOISE STATISTICS

With the VTS algorithm, we can estimate the clean feature vector sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ given a noisy feature vector sequence $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T\}$. It is assumed that the probability density function (PDF) of the clean feature vector is modeled by a mixture of Gaussian distributions such that

$$p(\mathbf{x}) = \sum_{k=1}^{M} p(k)\mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k) \tag{13}$$

where M is the total number of mixture components and $p(k)$, μ_k and Σ_k represent the given a priori probability, mean and variance of the k-th Gaussian distribution, respectively. We also assume that the statistics of the noise is represented by a single Gaussian distribution $\mathcal{N}(\mathbf{n}; \mu_{\mathbf{n}}, \Sigma_{\mathbf{n}})$ and $\mathbf{q}$ is an unknown constant vector. Under these assumptions, the environment compensation algorithm proceeds as follows [1]:

1. Get initial estimates for $\{\mu_{\mathbf{n}}, \Sigma_{\mathbf{n}}, \mathbf{q}\}$.
2. Using (2), expand $\mathbf{f}(\mathbf{n}, \mathbf{x}, \mathbf{q})$ around $(\mu_{\mathbf{n}}, \mu_k, \mathbf{q})$ for $k = 1, 2, \cdots, M$.
3. Perform a single iteration of the EM algorithm to reestimate the values of $\{\mu_{\mathbf{n}}, \Sigma_{\mathbf{n}}, \mathbf{q}\}$.
4. If the likelihood of the observed noisy data has not converged, return to Step 2.
5. Estimate the clean feature vector sequence based on the minimum mean squared error (MMSE) criterion.

Given $\lambda = \{\mu_{\mathbf{n}}, \Sigma_{\mathbf{n}}, \mathbf{q}\}$ and the clean feature vector distribution shown in (13), the PDF of the noisy feature is represented by

$$p(\mathbf{z}|\lambda) = \sum_{k=1}^{M} p(k)p(\mathbf{z}|k, \lambda) \tag{14}$$

$$= \sum_{k=1}^{M} p(k)\mathcal{N}(\mathbf{z}; \mu_{\mathbf{z},k}, \Sigma_{\mathbf{z},k}) \tag{15}$$

where

$$\mu_{\mathbf{z},k} = (\nabla_{\mathbf{x}}\mathbf{f})'\mu_k + (\nabla_{\mathbf{n}}\mathbf{f})'\mu_{\mathbf{n}} + (\nabla_{\mathbf{q}}\mathbf{f})'\mathbf{q} + \mathbf{g}(\mathbf{n}_0, \mathbf{x}_0, \mathbf{q}_0) \tag{16}$$

$$\Sigma_{\mathbf{z},k} = (\nabla_{\mathbf{x}}\mathbf{f})'\Sigma_k(\nabla_{\mathbf{x}}\mathbf{f}) + (\nabla_{\mathbf{n}}\mathbf{f})'\Sigma_{\mathbf{n}}(\nabla_{\mathbf{n}}\mathbf{f}) \ . \tag{17}$$

With (15), we can derive the a posteriori probability for each mixture component such that

$$p(k|\mathbf{z}, \lambda) = \frac{p(k)p(\mathbf{z}|k, \lambda)}{\sum_{k'=1}^{M} p(k')p(\mathbf{z}|k', \lambda)} \ . \tag{18}$$

The optimal parameter values $\lambda^o = \{\mu_{\mathbf{n}}^o, \Sigma_{\mathbf{n}}^o, \mathbf{q}^o\}$ are obtained according to the maximum likelihood (ML) criterion such that

$$\lambda^o = \underset{\lambda}{\mathrm{argmax}} \ [\log p(\mathbf{Z}|\lambda)] \ . \tag{19}$$

Since, however, it is difficult to directly obtain the ML estimates, the EM algorithm which iteratively updates the parameter values is adopted [4]. For two parameter sets $\lambda = \{\mu_{\mathbf{n}}, \Sigma_{\mathbf{n}}, \mathbf{q}\}$ and $\bar{\lambda} = \left\{\bar{\mu}_{\mathbf{n}}, \bar{\Sigma}_{\mathbf{n}}, \bar{\mathbf{q}}\right\}$, the auxiliary function, Q used in the EM algorithm is defined by

$$Q(\bar{\lambda}, \lambda) = E\left[\log p(\mathbf{X}, \mathbf{N}, \mathbf{K}|\lambda) \mid \mathbf{Z}, \bar{\lambda}\right] \tag{20}$$

where $\mathbf{N} = \{\mathbf{n}_1, \mathbf{n}_2, \cdots, \mathbf{n}_T\}$ denotes the noise vector sequence which is statistically independent of the clean feature vector sequence X, and $\mathbf{K} = \{k_1, k_2, \cdots, k_T\}$ is a hidden sequence of mixture components.

Let $\bar{\lambda} = \left\{\bar{\mu}_{\mathbf{n}}, \bar{\Sigma}_{\mathbf{n}}, \bar{\mathbf{q}}\right\}$ be the current parameter values. Then, the new parameter values $\hat{\lambda} = \left\{\hat{\mu}_{\mathbf{n}}, \hat{\Sigma}_{\mathbf{n}}, \hat{\mathbf{q}}\right\}$ are obtained according to

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \; Q(\bar{\lambda}, \lambda) \; . \tag{21}$$

By following the approach used in [2], we are led to

$$\hat{\mu}_{\mathbf{n}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{M} \mu_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda}) p(k|\mathbf{z}_t, \bar{\lambda}) \tag{22}$$

$$\hat{\Sigma}_{\mathbf{n}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{M} p(k|\mathbf{z}_t, \bar{\lambda}) \left[\Sigma_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda}) + \sigma^2_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda})\right] - \hat{\mu}_{\mathbf{n}}\hat{\mu}'_{\mathbf{n}} \tag{23}$$

$$\hat{\mathbf{q}} = \left[\sum_{t=1}^{T}\sum_{k=1}^{M} p(k|\mathbf{z}_t, \bar{\lambda})\Sigma^{-1}_{\mathbf{q},k}\right]^{-1} \left[\sum_{t=1}^{T}\sum_{k=1}^{M} p(k|\mathbf{z}_t, \bar{\lambda})\Sigma^{-1}_{\mathbf{q},k}\mathbf{q}(\mathbf{z}_t, k, \bar{\lambda})\right] \tag{24}$$

where

$$\mu_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda}) = E\left[\mathbf{n}_t|\mathbf{z}_t, k, \bar{\lambda}\right] \tag{25}$$

$$\sigma^2_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda}) = \mu_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda})\mu'_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda}) \tag{26}$$

$$\Sigma_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda}) = E\left[\mathbf{n}_t\mathbf{n}'_t|\mathbf{z}_t, k, \bar{\lambda}\right] - \sigma^2_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda}) \tag{27}$$

$$\Sigma^{-1}_{\mathbf{q},k} = (\nabla_{\mathbf{q}}\mathbf{f})(\nabla_{\mathbf{x}}\mathbf{f})^{-1}\Sigma_k^{-1}(\nabla_{\mathbf{x}}\mathbf{f})^{-t}(\nabla_{\mathbf{q}}\mathbf{f})' \tag{28}$$

and

$$\mathbf{q}(\mathbf{z}_t, k, \bar{\lambda}) = (\nabla_{\mathbf{q}}\mathbf{f})^{-t}\left(\mathbf{z}_t - (\nabla_{\mathbf{x}}\mathbf{f})'\mu_k - (\nabla_{\mathbf{n}}\mathbf{f})'\mu_{\mathbf{n}}(\mathbf{z}_t, k, \bar{\lambda}) - \mathbf{g}(\mathbf{n}_0, \mathbf{x}_0, \mathbf{q}_0)\right) \; . \tag{29}$$

After finding λ^o, the clean feature vector estimate, $\hat{\mathbf{x}}_t$ at time t can be computed based on the MMSE criterion as follows.

$$\hat{\mathbf{x}}_t = \sum_{k=1}^{M} p(k|\mathbf{z}_t, \lambda^o) E\left[\mathbf{x}_t|\mathbf{z}_t, \lambda^o\right] \; . \tag{30}$$

4. BAYESIAN PARAMETER ESTIMATION

In the VTS approach, the environmental parameters such as the noise mean and variance are usually estimated based on a small number of samples spanning a word or a sentence. This may cause the parameter estimates to deviate far from the true parameter values. One possible way to enhance the robustness in environmental parameter estimation is to use a priori knowledge of the parameters, which can be obtained by extracting some statistics from a large number of noisy samples. Let us assume that we are given the prior density of the parameters. Then, we can adopt the Bayesian parameter estimation scheme in the VTS approach. We follow the similar approach used in [5].

The conjugate prior density of $\mathcal{N}(\mathbf{n}; \mu_{\mathbf{n}}, \Sigma_{\mathbf{n}})$ is defined by a normal-Wishart density of the form

$$\varphi(\mu_{\mathbf{n}}, r_{\mathbf{n}}|\tau, m, \alpha, u) = |r_{\mathbf{n}}|^{\frac{(\alpha-N)}{2}} \exp\left[-\frac{\tau}{2}(\mu_{\mathbf{n}} - m)' r_{\mathbf{n}}(\mu_{\mathbf{n}} - m)\right] \exp\left[-\frac{1}{2}\mathrm{tr}(u r_{\mathbf{n}})\right] \tag{31}$$

where (τ, m, α, u) are the prior density parameters such that $\alpha > N - 1$, $\tau > 0$, and $r_{\mathbf{n}} = \Sigma_{\mathbf{n}}^{-1}$. Let

$$H(\lambda) = \log\left[\varphi(\mu_{\mathbf{n}}, r_{\mathbf{n}}|\tau, m, \alpha, u)\right] \; . \tag{32}$$

Then, (21) is replaced by

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \; \left[Q(\bar{\lambda}, \lambda) + H(\lambda)\right] \; . \tag{33}$$

By solving (33), we can get the Bayesian parameter reestimation formulas as follows.

$$\hat{\mu}_{\mathbf{n}} = \frac{\tau}{\tau + T} m + \frac{T}{\tau + T}\hat{\mu}_{\mathbf{n},\mathrm{ML}} \tag{34}$$

$$\hat{\Sigma}_{\mathbf{n}} = \frac{u + \tau(m - \hat{\mu}_{\mathbf{n}})(m - \hat{\mu}_{\mathbf{n}})'}{(\alpha - N) + T} + \frac{T}{(\alpha - N) + T}\hat{\Sigma}_{\mathbf{n},\mathrm{ML}} \tag{35}$$

where $(\hat{\mu}_{\mathbf{n},\mathrm{ML}}, \hat{\Sigma}_{\mathbf{n},\mathrm{ML}})$ are the ML estimates given by (22) and (23).

5. EXPERIMENTS

Performances of the proposed methods were evaluated with speaker-independent isolated word recognition experiments. The vocabulary consists of 75 Korean phonetically-balanced words. 20 male speakers uttered the words once to construct the database for training and evaluation. Utterances from 15 speakers constructed the training data and those from the other 5 speakers were used for evaluation. each utterance was low-pass filtered with a cut-off frequency of 4.5 kHz and digitized with a sampling rate of 16 kHz. A 18th-order mel-scaled log filterbank energy vector was extracted for every frame of 10 ms. By applying discrete cosine transform (DCT), a 12th-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.

Environment compensation was done in the log spectral domain except for the case of codeword-dependent cepstral normalization (CDCN) [3] which was carried out in the cepstral domain for the purpose of performance comparison. For each frame, the noisy feature vector which was represented by 18 mel-scaled log filterbank energies was transformed to a clean feature estimate. Clean speech features were modeled by a mixture of 128 Gaussian PDF's with diagonal covariance matrices. Table I shows the recognition results when artificial white Gaussian noises were were added to the clean speech waveform by varying the signal-to-noise ratio (SNR). VTS without noise statistics assumes that the added noise feature is an unknown constant vector while the proposed method assumes it to be statistically distributed according to a Gaussian PDF with unknown mean and variance. The recognition performance of the VTS approach without noise statistics was seriously degraded as the SNR became lowered. On the other hand, the proposed method which uses the noise statistics yielded better recognition performances over all the SNR ranges compared to those without noise statistics.

We could further improve the recognition performance by applying the Bayesian parameter reestimation scheme. A prior density for the noise mean and variance was defined for each SNR. Since each band of the log spectrum was assumed to be statistically independent, the prior density falls on a normal-gamma density and we could obtain the prior density parameter values from a set of noise data as in [6]. A remarkable improvement in recognition rate with the Bayesian approach was found when the SNR was low.

6. CONCLUSIONS

In this paper, we presented an exact EM formulations for the VTS approach with some specific noise statistics. Moreover, a Bayesian parameter reestimation scheme was proposed to incorporate a priori knowledge of the environmental parameters. From the recognition results, we could find that the proposed methods are effective for environment compensation especially when the SNR is low.

Table 1. Recognition results under various SNR conditions; recognition rate (%)

SNR (dB)	0	10	20	30
no processing	7.5	48.8	71.2	86.4
CDCN	33.3	66.7	79.7	85.1
VTS without noise statistics	20.3	73.1	87.5	91.2
VTS with noise statistics	46.9	80.0	88.0	91.2
Bayesian VTS	55.7	83.5	87.5	91.2

REFERENCES

[1] 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.

[2] N. S. Kim, D. Y. Kim and C. K. Un, "Environment compensation based on VTS with noise statistics," submitted to *IEEE Signal Processing Letters*.

[3] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition. Ph. D. Thesis, Carnegie Mellon University*, 1990.

[4] R. C. Rose, E. M. Hofstetter and D. A. Reynolds, "Integrated models of signal and background with applications to speaker identification in noise," *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 245-257, April 1994.

[5] J. -L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech, Audio Processing*, vol. 2, No. 2, pp. 291-298, April 1994.

[6] C. H. Lee, C. H. Lin and B. H. Juang, "A study on adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, No. 4, April 1991.