



# A Comparative Study of Noise Estimation Algorithms for VTS-Based Robust Speech Recognition

Yong Zhao, Biing-Hwang (Fred) Juang

Center for Signal and Image Processing, Georgia Institute of Technology

## Abstract

We conduct a comparative study to investigate two noise estimation approaches for robust speech recognition using vector Taylor series (VTS) developed in the past few years. The first approach, iterative root finding (IRF), directly differentiates the EM auxiliary function and approximates the root of the derivative function through recursive refinements. The second approach, twofold expectation maximization (TEM), estimates noise distributions by regarding them as hidden variables in a modified EM fashion. Mathematical derivations reveal the substantial connection between the two approaches. Two experiments are performed in evaluating the performance and convergence rate of the algorithms. The first is to fit a GMM model to artificially corrupted samples that are generated through Monte Carlo simulation. The second is to perform speech recognition on the Aurora 2 database.

**Index Terms:** Robust speech recognition, vector Taylor series, noise estimation

## 1. Introduction

In the past several years, vector Taylor series (VTS), which provides a linear approximation to relate the noisy speech and its clean counterpart, has been shown successful in various robust speech recognition methods, such as feature compensation [1], model adaptation [2] and noise adaptive training [3].

The success of the VTS adaptation relies on the accurate estimation of noise parameters, which are often formulated in an expectation maximization (EM) framework. Regardless its variant applications, the VTS noise estimation algorithm can be roughly classified into two categories. The first approach, referred to as iterative root finding (IRF) in this paper, is to directly differentiate the conventional EM auxiliary function and approximate the root of the resulting non-linear derivative function through recursive refinements. In [1], Moreno provided an EM framework to estimate both the means of noise and channel. In [4] [5], the IRF approach was generalized to incorporate the estimation of noise variance and the adaptation of dynamic features. The second approach, referred to as twofold expectation maximization (TEM) in this paper, follows the derivation proposed by Rose [6]. In this approach, clean speech and noise distributions are regarded as hidden variables in addition to HMM states and mixture components, and then EM iterations take place in both state-frame alignments and noise parameter re-estimation. In [7] [3], the TEM scheme was formulated for different applications.

In this paper, we conduct a comparative study to investigate the strength and weakness of these two different approaches. We believe that such comparison could benefit the understanding of VTS-based robust speech recognition techniques. Beyond a simple experiment-wise comparison, we reveal the substantial connection between the two noise estimation approaches. This connection illustrates the relations

between the two algorithms in performance and convergence properties, and indicates the trade-off and feasibility of switching these algorithms for a specific application.

Two experiments are performed in evaluating the performance and convergence rate of the above noise estimation algorithms. The first is to fit a GMM model to artificially corrupted samples that are generated through Monte Carlo simulation. The second is to perform speech recognition on the Aurora 2 database [8] of connected digits.

## 2. Vector Taylor series adaptation

Assuming that a clean speech signal  $x(t)$  is corrupted by both additive noise  $n(t)$  and convolutional distortion  $h(t)$ , the resulting noisy speech  $y(t)$  can be expressed as:

$$y(t) = x(t) * h(t) + n(t) \quad (1)$$

In the mel-cepstral domain, the noisy speech can be established as a nonlinear function of its clean counterpart [1] [2],

$$y = x + h + C \ln(1 + \exp(C^{-1}(n - x - h))) \equiv x + g(x, n, h) \quad (2)$$

where  $C$  is the discrete cosine transformation matrix, and  $x$ ,  $n$ ,  $h$ , and  $y$  denote the static feature vectors of the clean speech, additive noise, channel distortion, and noisy speech, respectively.

Assuming that  $x$  and  $n$  are independent and Gaussian distributed as  $\mathcal{N}(\mu_x, \Sigma_x)$  and  $\mathcal{N}(\mu_n, \Sigma_n)$  respectively, and  $h = \mu_h$  is constant, the distribution of  $y$  can be estimated by the first-order VTS expansion around  $\mu_x$ ,  $\mu_n$ , and  $\mu_h$ .

$$\mu_y = \mu_x + g(\mu_x, \mu_n, \mu_h) \quad (3)$$

$$\Sigma_y = G_x \Sigma_x G_x^T + G_n \Sigma_n G_n^T \quad (4)$$

where  $G_x$  and  $G_n$  are Jacobian matrices given by

$$G_x = C \text{diag} \left( \frac{1}{1 + \exp(C^{-1}(\mu_n - \mu_x - \mu_h))} \right) C^{-1} \quad (5)$$

$$G_n = I - G_x \quad (6)$$

For the delta portion of MFCC features, the following adaptation formulas have been used:

$$\mu_{\Delta y} = G_x \mu_{\Delta x} \quad (7)$$

$$\Sigma_{\Delta y} = G_x \Sigma_{\Delta x} G_x^T + G_n \Sigma_{\Delta n} G_n^T \quad (8)$$

The delta/delta portion of MFCC features takes a similar form.

Given a clean acoustic HMM set and an estimate of the noise parameters  $\theta = \{\mu_h, \mu_n, \Sigma_n, \Sigma_{\Delta n}, \Sigma_{\Delta \Delta n}\}$ , applying the above adaptation formulas to each Gaussian component of the models produces the corresponding noisy speech models. The generated HMM set matches the target noise environment and obtains an improved performance against the noisy speech. On the other hand, the noise parameters  $\theta$  may be re-estimated over the given utterance using an EM algorithm.

The conventional noise estimation procedure for the VTS adaptation is summarized as follows according to [5]:

1. For each utterance, initialize the additive noise parameters using the start and end frames, and set the channel mean vector to 0.
2. Transform the clean acoustic models using the adaptation formulas and decode the utterance.
3. Refine the noise estimate by maximizing the likelihood with respect to the assumed hypothesis.
4. Transform the models again, decode the utterance to obtain a final recognition transcription.

The steps described above include two decoding passes and one iteration of EM re-estimation. If the best performance was desired, a multiple-iteration EM of Step 3 and a multi-pass decoding loop between Steps 3 and 4 may be necessary.

In the remaining of the session, the two noise estimation approaches are discussed.

### 2.1. Iterative root finding

In the IRF approach, the auxiliary function for an utterance is defined as

$$Q(\theta, \hat{\theta}) = \sum_t \sum_{j,k \in \Omega} \gamma_{jk}(t) \log p(o_t | j, k, \hat{\theta}) \quad (9)$$

where  $\gamma_{jk}(t)$  denotes the posterior probability for the the  $k$ -th Gaussian in the  $j$ -th state of the HMM, and  $o_t$  denotes the complete noisy speech vector,  $o_t = [y_t^T, \Delta y_t^T, \Delta \Delta y_t^T]^T$ . The derivative of  $Q$  is a non-linear function of noise parameters, and does not have a closed solution. The IRF approach approximates the root of the derivative function through recursive refinements.

#### 2.1.1. Estimating noise and channel mean

The estimation of the noise and channel mean follows the ML formulation described in [5]. The key is that the first-order VTS may be used again to approximate the static noisy speech mean by expanding around an old estimate of noise and channel means

$$\hat{\mu}_{y,jk} \approx \mu_{y,jk} + G_{x,jk}(\hat{\mu}_h - \mu_h) + G_{n,jk}(\hat{\mu}_n - \mu_n) \quad (10)$$

To determine the noise mean that maximizes the auxiliary function, take the derivative of  $Q$  with respect to  $\hat{\mu}_n$ , and equate it to zero, obtaining

$$\frac{\partial Q}{\partial \hat{\mu}_n} = \sum_t \sum_{j,k \in \Omega} \gamma_{jk}(t) G_{n,jk}^T \Sigma_{y,jk}^{-1} [y_t - \hat{\mu}_{y,jk}] = 0 \quad (11)$$

Substituting the approximation of  $\hat{\mu}_{y,jk}$  (10) into (11) with  $\hat{\mu}_h = \mu_h$ , noise mean can be updated as

$$\hat{\mu}_n = \mu_n + \left[ \sum_{j,k \in \Omega} \gamma_{jk} G_{n,jk}^T \Sigma_{y,jk}^{-1} G_{n,jk} \right]^{-1} \sum_{j,k \in \Omega} G_{n,jk}^T \Sigma_{y,jk}^{-1} c_{y,jk} \quad (12)$$

where we define the following sufficient statistics

$$\gamma_{jk} = \sum_t \gamma_{jk}(t) \quad (13)$$

$$c_{y,jk} = \sum_t \gamma_{jk}(t) (y_t - \mu_{y,jk}) \quad (14)$$

Similarly, the channel mean is estimated as

$$\hat{\mu}_h = \mu_h + \left[ \sum_{j,k \in \Omega} \gamma_{jk} G_{x,jk}^T \Sigma_{y,jk}^{-1} G_{x,jk} \right]^{-1} \sum_{j,k \in \Omega} G_{x,jk}^T \Sigma_{y,jk}^{-1} c_{y,jk} \quad (15)$$

#### 2.1.2. Estimating noise variance

In the literature, several methods have been proposed to estimate noise variance, such as gradient descent method in [4], Newton's method in [5]. In [9], we presented a fixed point ap-

proach, summarized as below, to recursively approximate noise variance, promising better performance and less computational complexity.

Differentiate  $Q$  with respect to the static noise variance,

$$\frac{\partial Q}{\partial \hat{\Sigma}_n} = \frac{1}{2} \sum_{j,k \in \Omega} G_{n,jk}^T \hat{\Sigma}_{y,jk}^{-1} (S_{y,jk} - \gamma_{jk} \hat{\Sigma}_{y,jk}) \hat{\Sigma}_{y,jk}^{-1} G_{n,jk} \quad (16)$$

where we define the sufficient statistic

$$S_{y,jk} = \sum_t \gamma_{jk}(t) (y_t - \mu_{y,jk})(y_t - \mu_{y,jk})^T \quad (17)$$

The resulting derivative function (16) is a nonlinear function of noise variance, and has no closed solution. Nevertheless, the derivative function (16) can be thought of as a sum of rational functions, where the numerator (the central item of each summand) is first-order and the denominator (the second and fourth items) is a square of noise variance plus a clean speech model variance. By substituting the previous estimate of noise variance  $\Sigma_n$  into the denominator parts, we obtain a linear equation to update the noise variance,

$$\sum_{j,k \in \Omega} \gamma_{jk} A_{jk} \hat{\Sigma}_n A_{jk}^T = \sum_{j,k \in \Omega} B_{jk} \quad (18)$$

where the following notations are defined

$$A_{jk} = G_{n,jk}^T \Sigma_{y,jk}^{-1} G_{n,jk} \quad (19)$$

$$B_{jk} = G_{n,jk}^T \Sigma_{y,jk}^{-1} (S_{y,jk} - \gamma_{jk} G_{x,jk} \Sigma_{x,jk} G_{x,jk}^T) \Sigma_{y,jk}^{-1} G_{n,jk} \quad (20)$$

For the estimation of the dynamic noise variance, one needs to replace the static parameters with the corresponding dynamic parts.

### 2.2. Twofold expectation maximization

In the TEM approach, clean speech and noise distributions are regarded as hidden variables in addition to HMM states and mixture components. EM iterations take place in both state-frame alignments and noise parameter re-estimation. Specifically, for the static portion of MFCC features, the auxiliary function is defined as [6]

$$Q(\theta, \hat{\theta}) = \sum_t \sum_{j,k \in \Omega} \iint p(x_t, n_t, j, k | y_t, \theta) \log p(x_t, n_t | j, k, \hat{\theta}) dx_t dn_t \quad (21)$$

The re-estimation formulations described below are based on the method presented in [7] [3].

#### 2.2.1. Estimating noise mean and variance

Differentiating  $Q$  with respect to  $\hat{\mu}_n$  and  $\hat{\Sigma}_n$ , and equating it to zero yields [6]

$$\hat{\mu}_n = \frac{1}{T} \sum_t \sum_{j,k \in \Omega} \gamma_{jk}(t) \mu_{n|y,jk}(t) \quad (22)$$

$$\hat{\Sigma}_n = \frac{1}{T} \sum_t \sum_{j,k \in \Omega} \gamma_{jk}(t) \left[ \Sigma_{n|y,jk} + \mu_{n|y,jk}(t) \mu_{n|y,jk}^T(t) \right] - \hat{\mu}_n \hat{\mu}_n^T \quad (23)$$

where  $\mu_{n|y,jk}(t)$  and  $\Sigma_{n|y,jk}(t)$  denote the mean and variance of the instantaneous noise distribution  $p(n_t | y_t, j, k, \theta)$ , and can be obtained as follows:

$$\mu_{n|y,jk}(t) = \mu_n + \Sigma_{ny,jk} \Sigma_{y,jk}^{-1} (y_t - \mu_{y,jk}) \quad (24)$$

$$\Sigma_{n|y,jk} = \Sigma_n - \Sigma_{ny,jk} \Sigma_{y,jk}^{-1} \Sigma_{ny,jk} \quad (25)$$

$$\Sigma_{ny,jk} = \Sigma_{ny,jk}^T = G_{n,jk} \Sigma_n \quad (26)$$

The expression for the noise mean and variance estimate can be made more clear by substituting (24) and (25) into (22) and (23), and followed by some arithmetic manipulations,

$$\hat{\mu}_n = \mu_n + \frac{1}{T} \Sigma_n \left[ \sum_{j,k \in \Omega} G_{n,jk}^T \Sigma_{y,jk}^{-1} c_{y,jk} \right] \quad (27)$$

$$\hat{\Sigma}_n = \Sigma_n - (\hat{\mu}_n - \mu_n)(\hat{\mu}_n - \mu_n)^T + \frac{1}{T} \Sigma_n \left[ \sum_{j,k \in \Omega} G_{n,jk}^T \Sigma_{y,jk}^{-1} (S_{y,jk} - \gamma_{jk} \Sigma_{y,jk}) \Sigma_{y,jk}^{-1} G_{n,jk} \right] \Sigma_n \quad (28)$$

It is worth noting the similarities between the re-estimation equations of the TEM approach (27) and (28), and those of the IRF approach (12) and (18). Both approaches exhibit an update structure similar to that of the gradient descent method. The new estimate is basically obtained through the old estimate plus a correction term. For the noise mean estimation, the correction term is the gradient of  $Q$  at  $\mu_n$  (11) multiplied with a step-size-like term, the inverse of which is  $T \Sigma_n^{-1}$  in TEM and  $\sum_{j,k \in \Omega} \gamma_{jk} G_{n,jk}^T \Sigma_{y,jk}^{-1} G_{n,jk}$  in IRF, respectively. For noise variance, we observe a similar correspondence between the two approaches.

For the TEM approach, we note from (27) that if  $\Sigma_n$  is small, the convergence is slow. In an extreme case of initial noise variance being 0, there will be no update at all for both mean and variance. The relationship of the convergence to initial values will be examined in Section 3.

### 2.2.2. Estimating dynamic noise variance

For the estimation of the dynamic noise variance, one can think the dynamic noise, say delta noise, is Gaussian distributed with mean 0 and variance  $\Sigma_{\Delta n}$ . Similar to the derivation of (28), the delta noise variance is estimated as

$$\hat{\Sigma}_{\Delta n} = \Sigma_{\Delta n} + \frac{1}{T} \Sigma_{\Delta n} \left[ \sum_{j,k \in \Omega} G_{n,jk}^T \Sigma_{\Delta y,jk}^{-1} (S_{\Delta y,jk} - \gamma_{jk} \Sigma_{\Delta y,jk}) \Sigma_{\Delta y,jk}^{-1} G_{n,jk} \right] \Sigma_{\Delta n} \quad (29)$$

### 2.2.3. Estimating channel mean

Estimating channel mean in the TEM scheme is a tricky issue. The channel is assumed a deterministic quantity in a given utterance. If we estimate the channel mean in a straightforward way of TEM, the channel mean estimate does not change at all, as mentioned above. An alternate approach consistent with TEM is to presume that the clean speech feature in the noisy environment is distributed with mean  $\mu_{x,jk} + \mu_h$  and variance  $\Sigma_{x,jk}$ , where  $\mu_{x,jk}$  and  $\Sigma_{x,jk}$  are prior known. Differentiating  $Q$  (21) with respect to  $\hat{\mu}_h$  and equating it to zero gives

$$\hat{\mu}_h = \mu_h + \left[ \sum_{j,k \in \Omega} \gamma_{jk} \Sigma_{x,jk}^{-1} \right]^{-1} \sum_{j,k \in \Omega} G_{x,jk}^T \Sigma_{y,jk}^{-1} c_{y,jk} \quad (30)$$

The above channel re-estimation formula is essentially the same as the one proposed in [7], but gives a more intuitive derivation. By comparing (30) and (15), we see again the connection between the two noise estimation approaches.

## 3. Simulation on a GMM fitting task

In the following experiments, three noise estimation methods are compared, namely, *IRF-1*: the IRF approach depicted in Subsection 2.1, i.e., noise and channel mean are estimated using (12) and (15), and noise variances are estimated using the fixed point method (18); *IRF-2*: noise variances are estimated using Newton's method as described in [5], others being the same as IRF-1; *TEM*: the TEM approach depicted in Subsection 2.2.

In this section, the quality of the noise estimation algorithms has been investigated in a GMM fitting task, where the test data are generated through Monte Carlo simulation. For the simplicity of analysis, the distortion model (2) is considered in the log-spectrum domain without convolutional noise,

$$y = x + \ln(1 + \exp(n - x)) \quad (31)$$

we first artificially generate the clean signal vectors, which consist of 8 Gaussian components of 8-dimensional data. 125 samples are drawn from each of the 8 components, and there are 1000 samples in total. Mean values of these components at each dimension are randomly chosen from -20 to 20, and variances from 1/4 to 16. The clear signals are then contaminated through (31) with noise data that are drawn from a Gaussian distribution with mean 0 and variance 4 at each dimension.

We run 10 trials of simulations. For each trial, the noise estimation algorithms are repeated with different initial noise means and variances. The initial means range from -2 to 2 at step size 0.5, and the initial variances range from 1/8 to 32 increasing by a multiplication with 2. The EM iterations are stopped whenever the auxiliary function fails to increase by a certain threshold (0.1%).

Table 1 shows the comparison of accuracies and convergence rates of the three algorithms. We note that though these algorithms achieve the same level of accuracy, the IRF methods converge significantly faster than the TEM method, and also less sensitive to the variation of the initial noise settings.

Table 1: Comparison of convergence rate and algorithmic accuracies of the noise estimation algorithms by fitting a GMM model to artificially corrupted samples.

	# of iterations		Log-likelihood per sample
	mean	std. dev.	
TEM	9.24	5.94	-17.99
IRF-1	2.29	0.74	-17.97
IRF-2	2.96	1.11	-17.99

To examine more carefully the effect of initial noise values to the TEM method, Fig. 1 illustrates the variation of the number of iterations with different initial values. It is observed that the convergence depends on not only initial noise variances but also initial noise means. In the neighborhood of reference noise values, the change of iteration numbers is roughly proportional to the change of initial means and the change of the logarithm of initial variances. When both initial noise means and noise variances are small, the convergence becomes substantially slow.

## 4. Recognition experiments and results

In this experiment, the noise estimation algorithms were evaluated on the Aurora 2 database [8] of connected digits. The test set consists of three different parts. Test Set A and Test Set B each contain 4 types of additive noises, and the data in Test Set C are contaminated with 2 types of additive noises as well as channel distortion. For each noise type, a subset of the clean speech utterances is contaminated at SNRs ranging from 20 to -5 dB at a 5 dB step size, which, including the clean condition, constitute 7 different SNR levels.

The clean training set is used to estimate the baseline ML HMMs. The acoustic models were trained using the standard Aurora 2 recipe for the simple back end. 39-dimensional MFCC features with the 0th cepstral coefficient for the energy term are used in the experiments. The cepstra are computed based on spectral magnitude. The baseline ML system yields word error

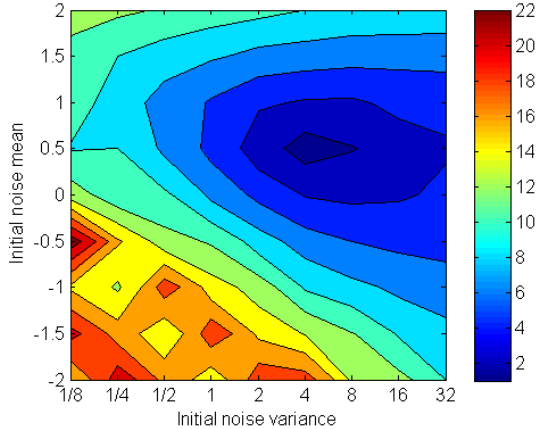


Figure 1: Variation of the number of iterations with respect to different initial values for the TEM method in a GMM fitting task.

rate (WER) of 41.57% by averaging over SNRs between 20 and 0 dB of three test sets.

We particularly perform VTS noise adaptation experiments using two methods, IRF-1 and TEM. The noise re-estimation procedure is summarized in Section 2. The first and last 20 frames of each utterance, assumed from the non-speech area, are used for initializing the means and variances of additive noise. The channel is initialized to 0. The speech models transformed using the initial noise estimates achieve 12.86% WER.

Fig. 2 shows the performance comparison as a function of the total number of re-estimation iterations, which is calculated as

$$\text{total \# of re-est} = (\# \text{ of decoding passes} - 1) \times (\# \text{ of re-est/pass})$$

Note that the first decoding pass is based on the transformed models that take initial noise estimates. It observes that the two approaches achieve the similar performance given sufficient iteration steps, but the IRF method converges significantly faster than the TEM method. The figure also shows that additional decoding passes, which interleaves between the sequence of re-estimations, do not considerably benefit to the estimation of the noise parameters.

Table 2: Performance and convergence properties of the noise estimation algorithms in two stopping criteria.

	Stop at min. WER		Stop at 1% WER reduction	
	WER (%)	# of re-est.	WER (%)	# of re-est.
TEM	8.32	96	9.24	8
IRF-1	8.31	4	8.42	2

Table 2 tabulates the recognition performance and convergence properties of the noise estimation algorithms in two stopping criteria. The first case is the minimum WER that the algorithms can achieve in a long term. The second case compares the performance in a more practical setup, i.e., we limit the algorithms to two decoding passes and stop EM iterations when WER fails to decrease by a certain threshold (1%). Apparently, the IRF approach achieves a significantly better performance in both recognition accuracy and computational complexity in comparison with the TEM approach.

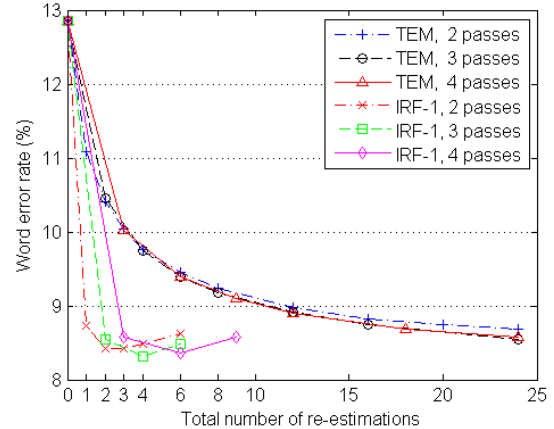


Figure 2: Performance (WER in %) comparison as a function of the total number of re-estimation iterations.

## 5. Conclusion

We compared two mainstream noise estimation approaches for robust speech recognition using VTS. Mathematical derivations revealed the substantial connection between the two approaches. Experimental results demonstrated the advantage of the IRF approach over the TEM approach.

It does not mean that the TEM approach is less favorable than the IRF approach. When we go beyond VTS-based robust speech recognition, where noisy speech statistics need not be a closed function of noise statistics, the TEM approach have to be considered. In the TEM approach, the essential quantities required to correlate noisy speech statistics and noise statistics are  $\Sigma_{yn}$ .

## 6. References

- [1] P. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, 1996.
- [2] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *ICSLP*, Beijing, China, 2000, pp. 869–872.
- [3] Y. Hu and Q. Huo, "Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions," in *INTERSPEECH*, Antwerp, Belgium, 2007.
- [4] H. Liao, *Uncertainty Decoding for Noise Robust Speech Recognition*, Ph.D. thesis, University of Cambridge, 2007.
- [5] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer Speech and Language*, vol. 23, pp. 389–405, 2009.
- [6] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, 1994.
- [7] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first order vector Taylor series," *Speech Communication*, vol. 24, pp. 39–49, 1998.
- [8] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR*, 2000.
- [9] Y. Zhao and B.-H. Juang, "On noise estimation for robust speech recognition using vector Taylor series," in *ICASSP*, Dallas, TX, 2010.