interspeech 2008 incorporating SST 08

A Feature Compensation Approach Using High-Order Vector Taylor Series Approximation of an Explicit Distortion Model for Noisy Speech Recognition

Jun Du[1,2], *Qiang Huo*[1]

[1]Microsoft Research Asia, Beijing, P. R. China
[2]University of Science and Technology of China, Hefei, Anhui, P. R. China
unuedjwj@ustc.edu, qianghuo@microsoft.com

Abstract

This paper presents a new feature compensation approach to noisy speech recognition by using high-order vector Taylor series (HOVTS) approximation of an explicit model of environmental distortions. Formulations for maximum likelihood (ML) estimation of noise model parameters and minimum mean-squared error (MMSE) estimation of clean speech are derived. Experimental results on Aurora2 database demonstrate that the proposed approach achieves consistently significant improvement in recognition accuracy compared to traditional first-order VTS based feature compensation approach.

Index Terms— robust speech recognition, feature compensation, vector Taylor series, distortion model.

1. Introduction

Most of current automatic speech recognition (ASR) systems use MFCCs (Mel-Frequency Cepstral Coefficients) and their derivatives as speech features, and a set of Gaussian mixture continuous density HMMs (CDHMMs) for modeling basic speech units. It is well known that the performance of such an ASR system trained with clean speech will degrade significantly when the testing speech is distorted by additive noises. How to achieve the noise robustness has been an important research topic in ASR field. Among many approaches proposed previously, one type of approaches is the so-called feature compensation approach using *explicit* model of environmental distortions (e.g., [5, 4]), which is also the topic of this paper. For our approach, it is assumed that in the time domain, the "corrupted" speech $y[t]$ is subject to the following *explicit* distortion model:

$$y[t] = x[t] + n[t] \tag{1}$$

where independent signals $x[t]$ and $n[t]$ represent the t^{th} sample of clean speech and additive noise, respectively. By ignoring correlations between different filter banks, the distortion model in log power-spectrum domain can be expressed *approximately* as

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

where $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$ are log power-spectrums in a particular channel of the filterbank of noisy speech, clean speech and noise, respectively. The nonlinear nature of the above distortion model makes statistical modeling and inference of the above variables difficult, therefore certain approximations have to be made. Understandably, a simple linear approximation, namely the first-order vector Taylor series (VTS) approximation, has been tried in the past (e.g., [5, 4]).

This work has been done when the first author was an intern at Microsoft Research Asia, Beijing, China.

There are also efforts in using high-order VTS (HOVTS) to improve the above first-order VTS approximation. In [3], the above nonlinear distortion function is first expanded using HOVTS. Then a linear function is found to approximate the above HOVTS by minimizing the mean-squared error incurred by this approximation. Given the linear function, the remaining inference is the same as in using the traditional first-order VTS to approximate the nonlinear distortion function directly. In [7], the above nonlinear distortion function is approximated by a second-order VTS. Using this relation, the mean vector of the relevant noisy speech feature vector can be derived, which naturally includes a term related to the second-order term in HOVTS. In this paper, we extend the work in [7] in the following ways: 1) the nonlinear distortion function can be approximated by HOVTS with any order, 2) the required sufficient statistics are derived for estimating both noise model parameters and clean speech feature vector. In comparison with the work in [3], correlations among different channels of filterbank can be considered by our approach.

The rest of the paper is organized as follows. In Section 2, we give an overview of the general formulation of our feature compensation approach. In Section 3, we present the detailed formulation of how to calculate the required sufficient statistics based on HOVTS approximation. In Section 4, we report experimental results and finally we conclude the paper in Section 5.

2. Our Feature Compensation Approach

The flowchart of our feature compensation approach is illustrated in Fig. 1. In the training stage, a Gaussian mixture model (GMM), $p(\mathbf{x}_t^c) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\mathbf{x}_t^c; \boldsymbol{\mu}_{\mathbf{x},m}^c, \boldsymbol{\Sigma}_{\mathbf{x},m}^c)$, is trained from clean speech using MFCC features without cepstral mean normalization (CMN), where $\boldsymbol{\mu}_{\mathbf{x},m}^c$, $\boldsymbol{\Sigma}_{\mathbf{x},m}^c$, and ω_m are mean vector, diagonal covariance matrix and mixture weight of the m^{th} component, respectively. The relevant model parameters can be transformed from cepstral domain to log-power-spectral domain for later use as follows:

$$\boldsymbol{\mu}_{\mathbf{x},m}^{\mathrm{l}} = \mathbf{C}^{+}\boldsymbol{\mu}_{\mathbf{x},m}^{\mathrm{c}} \tag{3}$$

$$\boldsymbol{\Sigma}_{\mathbf{x},m}^{\mathrm{l}} = \mathbf{C}^{+}\boldsymbol{\Sigma}_{\mathbf{x},m}^{\mathrm{c}}(\mathbf{C}^{+})^{\top} \tag{4}$$

where $\mathbf{C}^{+}$ is the Moore-Penrose inverse [2] of the discrete cosine transform (DCT) matrix $\mathbf{C}$, the superscript 'l' and 'c' indicate the log-power-spectral domain and cepstral domain, respectively.

Let's assume that for each sentence, the noise feature vector $\mathbf{n}^c$ in cepstral domain follows a Gaussian PDF (probability density function) with a mean vector $\boldsymbol{\mu}_{\mathbf{n}}^c$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_{\mathbf{n}}^c$, which can be estimated in the recognition stage as follows:

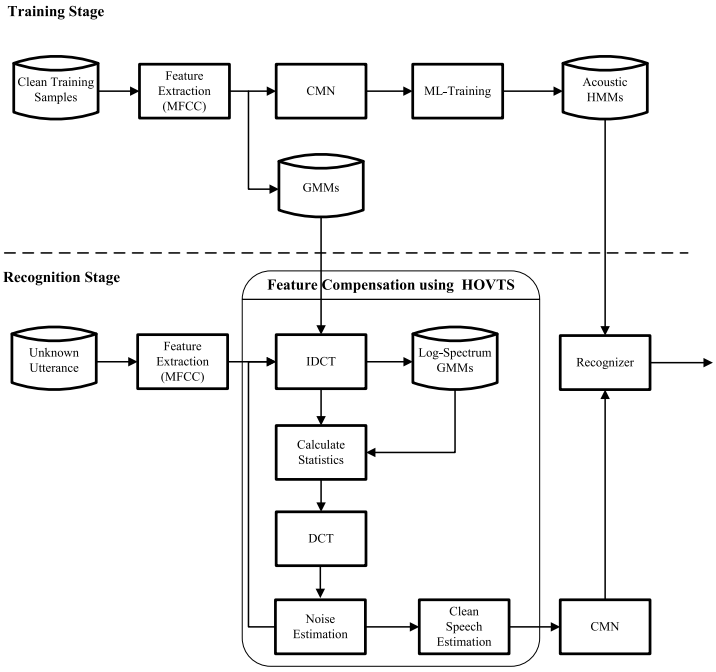

Figure 1: Flowchart of our feature compensation approach.

Step 1: Initialization:
We first estimate the initial noise model parameters in cepstral domain by simply taking the sample mean and covariance of the MFCC features from the first several (10 in our experiments) frames of the unknown utterance.

Step 2: Transform noise model parameters from cepstral domain to log-power-spectral domain as follows:

$$\boldsymbol{\mu}_{\mathbf{n}}^{\mathrm{l}} = \mathbf{C}^{+}\boldsymbol{\mu}_{\mathbf{n}}^{\mathrm{c}} \quad (5)$$

$$\boldsymbol{\Sigma}_{\mathbf{n}}^{\mathrm{l}} = \mathbf{C}^{+}\boldsymbol{\Sigma}_{\mathbf{n}}^{\mathrm{c}}(\mathbf{C}^{+})^{\top} . \quad (6)$$

Step 3: In log-power-spectral domain, use HOVTS approximation to calculate, as described in next section, the relevant statistics, $\boldsymbol{\mu}_{\mathbf{y},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{y},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{xy},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{ny},m}^{\mathrm{l}}$, which are required for noise re-estimation and clean speech estimation.

Step 4: Transform the above statistics back to cepstral domain as follows:

$$\boldsymbol{\mu}_{\mathbf{y},m}^{\mathrm{c}} = \mathbf{C}\boldsymbol{\mu}_{\mathbf{y},m}^{\mathrm{l}} \quad (7)$$

$$\boldsymbol{\Sigma}_{\mathbf{y},m}^{\mathrm{c}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{y},m}^{\mathrm{l}}(\mathbf{C})^{\top} \quad (8)$$

$$\boldsymbol{\Sigma}_{\mathbf{xy},m}^{\mathrm{c}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{xy},m}^{\mathrm{l}}(\mathbf{C})^{\top} \quad (9)$$

$$\boldsymbol{\Sigma}_{\mathbf{ny},m}^{\mathrm{c}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{ny},m}^{\mathrm{l}}(\mathbf{C})^{\top} . \quad (10)$$

Step 5: Use the following updating formulas (e.g., [6, 4]) to re-estimate the noise model parameters:

$$\overline{\boldsymbol{\mu}}_{\mathbf{n}} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t,m]}{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)} \quad (11)$$

$$\overline{\boldsymbol{\Sigma}}_{\mathbf{n}} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)E_{\mathbf{n}}[\mathbf{n}_t\mathbf{n}_t^{\top}|\mathbf{y}_t,m]}{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)} - \overline{\boldsymbol{\mu}}_{\mathbf{n}}\overline{\boldsymbol{\mu}}_{\mathbf{n}}^{\top} \quad (12)$$

where

$$P(m|\mathbf{y}_t) = \frac{\omega_m p_{\mathbf{y}}(\mathbf{y}_t|m)}{\sum_{l=1}^{M}\omega_l p_{\mathbf{y}}(\mathbf{y}_t|l)} . \quad (13)$$

In the above equations, we have dropped the cepstral domain indicator "c" in relevant variables for notational convenience. Furthermore, $p_{\mathbf{y}}(\mathbf{y}_t|m)$ is the PDF of the noisy speech $\mathbf{y}_t$ for the m^{th} component of the mixture of densities for the compensated noisy speech, which is approximated by a Gaussian PDF, $\mathcal{N}(\mathbf{y}_t;\boldsymbol{\mu}_{\mathbf{y},m},\boldsymbol{\Sigma}_{\mathbf{y},m})$, via "moment-matching". $E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t,m]$ and $E_{\mathbf{n}}[\mathbf{n}_t\mathbf{n}_t^{\top}|\mathbf{y}_t,m]$ are the relevant conditional expectations evaluated as follows:

$$E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t,m] = \boldsymbol{\mu}_{\mathbf{n}} + \boldsymbol{\Sigma}_{\mathbf{ny},m}\boldsymbol{\Sigma}_{\mathbf{y},m}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},m}) \quad (14)$$

$$E_{\mathbf{n}}[\mathbf{n}_t\mathbf{n}_t^{\top}|\mathbf{y}_t,m] = E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t,m]E_{\mathbf{n}}^{\top}[\mathbf{n}_t|\mathbf{y}_t,m] + \boldsymbol{\Sigma}_{\mathbf{n}} - \boldsymbol{\Sigma}_{\mathbf{ny},m}\boldsymbol{\Sigma}_{\mathbf{y},m}^{-1}\boldsymbol{\Sigma}_{\mathbf{yn},m} . \quad (15)$$

Step 6: Repeat Step 2 to Step 5 several times.

Given the noisy speech and noise estimation, the minimum mean-squared error (MMSE) estimation of clean speech feature vector in cepstral domain can be calculated as

$$\hat{\mathbf{x}}_t = E_{\mathbf{x}}[\mathbf{x}_t|\mathbf{y}_t] = \sum_{m=1}^{M} P(m|\mathbf{y}_t)E_{\mathbf{x}}[\mathbf{x}_t|\mathbf{y}_t,m] \quad (16)$$

where $E_{\mathbf{x}}[\mathbf{x}_t|\mathbf{y}_t,m]$ is the conditional expectation of $\mathbf{x}_t$ given $\mathbf{y}_t$ for the m^{th} mixture component and can be evaluated as follows:

$$E_{\mathbf{x}}[\mathbf{x}_t|\mathbf{y}_t,m] = \boldsymbol{\mu}_{\mathbf{x},m} + \boldsymbol{\Sigma}_{\mathbf{xy},m}\boldsymbol{\Sigma}_{\mathbf{y},m}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},m}) . \quad (17)$$

The other modules in Fig. 1 are self-explained.

In next section, we elaborate on how to calculate the required statistics, $\boldsymbol{\mu}_{\mathbf{y},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{y},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{xy},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{ny},m}^{\mathrm{l}}$, using HOVTS approximation of the nonlinear distortion function in Eq. (2). For notational convenience, we drop hereinafter the indices related to the frame number, mixture component, and channel index of the filterbank without causing confusions.

3. Computation of Required Statistics

The explicit distortion model in Eq. (2) is reformulated in the scalar form as follows:

$$y = f(x,n) = \log(\exp(x) + \exp(n)) . \quad (18)$$

Then the K-order Taylor series of $f(x,n)$ with the expansion point (μ_x,μ_n) can be represented as

$$\begin{aligned} & f_K(x,n) \\ = & \sum_{k=0}^{K}\frac{1}{k!}\left[(x-\mu_x)\frac{\partial}{\partial x} + (n-\mu_n)\frac{\partial}{\partial n}\right]^k f(\mu_x,\mu_n) \\ = & \sum_{k=0}^{K}\sum_{r=0}^{k} A(k,r)(x-\mu_x)^{k-r}(n-\mu_n)^r \end{aligned} \quad (19)$$

where

$$A(k,r) = \frac{1}{r!(k-r)!}\left.\frac{\partial^k f(x,n)}{\partial x^{k-r}\partial n^r}\right|_{(\mu_x,\mu_n)} \quad (20)$$

and

$$\left.\frac{\partial^k f(x,n)}{\partial x^{k-r}\partial n^r}\right|_{(\mu_x,\mu_n)} = \begin{cases} \log(\exp(\mu_x)+\exp(\mu_n)), & k=0, r=0 \\ 1-\frac{1}{1+\exp(\mu_n-\mu_x)}, & k=1, r=1 \\ \frac{1}{1+\exp(\mu_n-\mu_x)}, & k=1, r=0 \\ (-1)^{k-r}\sum_{p=1}^{k}\frac{B(k,p)}{[1+\exp(\mu_n-\mu_x)]^p}, & k>1 \end{cases} . \quad (21)$$

When $k > 1$ and $k \geq p \geq 1$, the coefficients $B(k,p)$ in Eq. (21) can be evaluated by using the following recursive relation

$$B(k,p) = (p-1)B(k-1,p-1) - pB(k-1,p) \quad (22)$$

with the initial condition

$$B(1,1) = -1, B(k,0) = B(k,k+1) = 0, \ \ k \geq 1 \ . \quad (23)$$

For convenience, we also define the following expectations:

$$E^i_{xn}[g(x,n)] = \iint g(x^i,n^i)p_{xn}(x^i,n^i)dx^i dn^i \quad (24)$$

$$E^{ij}_{xn}[g(x,n),h(x,n)] = \iiiint g(x^i,n^i)h(x^j,n^j) p_{xn}(x^i,x^j,n^i,n^j)dx^i dx^j dn^i dn^j \quad (25)$$

where $g(x^i,n^i)$ and $h(x^j,n^j)$ are two general functions, i and j are dimensional indices.

Given the above notations and results, we summarize in the following subsections the main statistics required in implementing our feature compensation approach.

3.1. Calculating $\mu_y(i)$

Let's use $\mu_y(i)$ to denote the i^{th} element of the vector $\boldsymbol{\mu}_\mathbf{y}$. Using the definition of the mean parameter, we have

$$\begin{aligned} & \mu_y(i) \\ \doteq\ & E^i_{xn}[f_K(x,n)] \\ =\ & \sum_{k=0}^{K}\sum_{r=0}^{k} A^i(k,r)E^i_{xn}[(x-\mu_x)^{k-r}(n-\mu_n)^r] \\ =\ & \sum_{k=0}^{K}\sum_{r=0}^{k} A^i(k,r)M^i_n(r)M^i_x(k-r) \end{aligned} \quad (26)$$

where

$$M^i_\Delta(p) = \begin{cases} 0, & \text{if } p \text{ is odd} \\ (p-1)!!\sigma^p_\Delta(i), & \text{otherwise} \end{cases} \quad (27)$$

Δ represents 'x' or 'n'. $A^i(k,r)$ is the value of Eq. (20) for the i^{th} dimension.

3.2. Calculating $\sigma^2_y(i,j)$

Let's use $\sigma^2_y(i,j)$ to denote the $(i,j)^{\text{th}}$ element of the matrix $\boldsymbol{\Sigma}_\mathbf{y}$. Using the definition of the covariance, we have

$$\begin{aligned} & \sigma^2_y(i,j) \\ \doteq\ & E^{ij}_{xn}[f_K(x,n),f_K(x,n)] - \mu_y(i)\mu_y(j) \\ =\ & \sum_{k_1=0}^{K}\sum_{r_1=0}^{k_1}\sum_{k_2=0}^{K}\sum_{r_2=0}^{k_2} A^i(k_1,r_1)A^j(k_2,r_2)M^{ij}_n(r_1,r_2) \\ & M^{ij}_x(k_1-r_1,k_2-r_2) - \mu_y(i)\mu_y(j) \end{aligned} \quad (28)$$

where

$$M^{ij}_\Delta(p,q) = \begin{cases} 0, & \text{if } p+q \text{ is odd} \\ p!q!2^{-\frac{p+q}{2}}\sum_{0\leq l\leq min(p,q)}^{p-l \text{ is even}} & \\ \frac{2^l}{l!(\frac{p-l}{2})!(\frac{q-l}{2})!}\sigma^{p-l}_\Delta(i,i) & \\ \sigma^{2l}_\Delta(i,j)\sigma^{q-l}_\Delta(j,j), & \text{otherwise} \end{cases} . \quad (29)$$

3.3. Calculating $\sigma^2_{xy}(i,j)$

Let's use $\sigma^2_{xy}(i,j)$ to denote the $(i,j)^{\text{th}}$ element of the matrix $\boldsymbol{\Sigma}_{\mathbf{xy}}$. Using the definition of the covariance parameter, we have

$$\begin{aligned} & \sigma^2_{xy}(i,j) \\ =\ & E^{ij}_{xn}[(x-\mu_x),(y-\mu_y)] \\ =\ & \sum_{k=0}^{K}\sum_{r=0}^{k} A^j(k,r)M^j_n(r)M^{ij}_x(1,k-r) \ . \end{aligned} \quad (30)$$

3.4. Calculating $\sigma^2_{ny}(i,j)$

Let's use $\sigma^2_{ny}(i,j)$ to denote the $(i,j)^{\text{th}}$ element of the matrix $\boldsymbol{\Sigma}_{\mathbf{ny}}$. Using the definition of the covariance parameter, we have

$$\begin{aligned} & \sigma^2_{ny}(i,j) \\ =\ & E^{ij}_{xn}[(n-\mu_n),(y-\mu_y)] \\ =\ & \sum_{k=0}^{K}\sum_{r=0}^{k} A^j(k,r)M^{ij}_n(1,r)M^j_x(k-r) \ . \end{aligned} \quad (31)$$

4. Experiments and Results

4.1. Experimental Setup

In order to verify the effectiveness of the proposed approach, a series of experiments are performed for the task of speaker independent recognition of connected digit strings on Aurora2 database. A full description of the Aurora2 database and a test framework is given in [1].

In our ASR systems, the feature vector we used consists of 13 MFCCs (including C_0) plus their first and second order derivatives. The number of Mel-frequency filter banks is 23. MFCCs are computed based on power spectrum. Each digit is modeled by a whole-word left-to-right CDHMM, which consists of 16 emitting states, each having 3 Gaussian mixture components. The mixture number of clean-speech GMM for feature compensation is 256. "Clean-training" is used. Our baseline systems refer to the ones with CMN but no other feature compensation applied.

4.2. Experimental Results

Table 1 summarizes a performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using HOVTS-based feature compensation for cases of using first 10 frames to estimate noise model parameters (referred to as "NO"), and using ML noise re-estimation (four EM iterations, referred to as "YES"). The performance is averaged over SNRs between 0dB and 20dB on test Set A, Set B and Set C respectively. Several observations can be made. First, all the robust systems using HOVTS-based feature compensation outperform the "Baseline" system. Second, all the feature compensation methods using noise re-estimation perform better than those without noise re-estimation. Third, "VTS(3)" performs better than "VTS(2)", and "VTS(2)" outperforms "VTS(1)".

Table 3: Detailed results of VTS(3)-based feature compensation using noise reestimation on Aurora2 database.

Clean Training - Results											
	Set A				Set B				Set C		
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Subway M	Street M	Avg.
Clean	98.93	99.21	99.05	99.07	98.93	99.21	99.05	99.07	98.96	99.24	99.07
20dB	98.04	98.34	98.78	98.49	98.53	97.91	98.72	98.70	98.07	97.91	98.35
15dB	96.41	97.22	98.00	96.85	97.67	96.43	97.88	97.81	95.98	96.58	97.08
10dB	93.28	93.86	95.74	93.61	94.44	93.20	95.74	95.09	92.08	92.14	93.92
5dB	85.78	83.25	88.55	84.23	82.87	83.86	87.35	86.24	83.24	80.96	84.63
0dB	65.89	56.29	66.00	64.52	59.23	61.46	68.33	64.52	61.38	53.57	62.12
-5dB	31.78	21.49	25.08	32.49	24.99	26.27	31.05	28.17	28.15	23.40	27.29
Avg.	87.88	85.79	89.41	87.54	86.55	86.57	89.60	88.47	86.15	84.23	87.22

Table 1: Performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using HOVTS-based feature compensation, averaged over SNRs between 0dB and 20dB across all noise conditions on three different test sets of Aurora2 database.

Methods		Set A	Set B	Set C	Overall
Baseline		66.55	71.57	67.43	68.74
NO	VTS(1)	85.10	85.38	82.80	84.75
	VTS(2)	85.77	86.27	83.70	85.55
	VTS(3)	86.69	86.90	84.84	86.41
YES	VTS(1)	86.24	86.52	83.86	85.88
	VTS(2)	86.90	87.19	84.57	86.55
	VTS(3)	87.66	87.80	85.19	87.22

Table 2: Performance (word accuracy in %) comparison of several methods averaged over three test sets of Aurora2 dababase at each SNR.

Methods		0dB	5dB	10dB	15dB	20dB
Baseline		24.64	49.33	79.57	93.04	97.10
NO	VTS(1)	56.92	80.92	91.95	96.20	97.78
	VTS(2)	57.89	82.10	92.91	96.74	98.14
	VTS(3)	59.88	83.56	93.47	96.94	98.18
YES	VTS(1)	58.83	82.82	92.99	96.72	98.03
	VTS(2)	60.33	83.67	93.56	96.92	98.28
	VTS(3)	62.12	84.63	93.92	97.08	98.35

Apparently the third-order information seems useful, but no further improvement is observed when the order is more than three.

Table 2 gives a performance (word accuracy in %) comparison of several methods averaged over three test sets of Aurora2 database at each SNR (in dB). Similar observations can also be made under different SNRs.

Overall, "VTS(3)" using noise re-estimation achieves the best performance in all testing conditions. Detailed results for this approach are listed in Table 3.

5. Summary and Future Work

In this paper, we have proposed a feature compensation approach using high-order vector Taylor series (HOVTS) approximation of an explicit distortion model. Its effectiveness has been confirmed in an experimental study on both Aurora2 and Aurora3 databases, but only experimental results on Aurora2 are reported in this paper due to the page limit. Ongoing and future works include

- to study HOVTS-based HMM compensation,
- to explore irrelevant variability normalization (IVN) based HMM training using HOVTS,
- to apply the similar idea to speech enhancement.

We will report those results elsewhere when they become available.

6. References

[1] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *Proc. ISCA ITRW ASR*, 2000, pp.181-188.

[2] Y. Hu and Q. Huo, "Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions," *Proc. Interspeech*, 2007, pp.1042-1045.

[3] N. S. Kim, "Statistical linear approximation for environment compensation," *IEEE Signal Processing Letters*, Vol. 5, No. 1, pp.8-10, 1998.

[4] 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.

[5] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," *Proc. ICASSP*, 1996, pp.733-736.

[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 Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp.245-257, 1994.

[7] V. Stouten, *Robust Automatic Speech Recognition in Time-Varying Environemnts*, Ph.D. thesis, Katholieke Universiteit Leuven, 2006.

[8] G.-H. Ding, B. Xu, "Exploring high-performance speech recognition in noisy environments using high-order Taylor series expansion," *Proc. ICSLP*, 2004, pp.149-152.