EVALUATION OF A FEATURE COMPENSATION APPROACH USING HIGH-ORDER VECTOR TAYLOR SERIES APPROXIMATION OF AN EXPLICIT DISTORTION MODEL ON AURORA2, AURORA3, AND AURORA4 TASKS

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

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

ABSTRACT

In our previous work, a new feature compensation approach to robust speech recognition was proposed by using high-order vector Taylor series (HOVTS) approximation of an explicit model of distortions caused by additive noises, and evaluation results were reported on Aurora2 database. This paper extends the above approach to deal with both additive noises and convolutional distortions, and reports evaluation results on Aurora2, Aurora3, and Aurora4 tasks.

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 corrupted by additive noises and convolutional distortions. One type of approaches to deal with the above problem is the so-called feature compensation approach using *explicit* model of environmental distortions (e.g., [13, 12]), 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:

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

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

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

where $\mathbf{y}$, $\mathbf{x}$, $\mathbf{h}$ and $\mathbf{n}$ are log power-spectrums in a particular channel of the filterbank of noisy speech, clean speech, convolutional term 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., [13, 12]). There are also efforts in using high-order VTS (HOVTS) to improve the above first-order VTS approximation. In [11], the nonlinear distortion function for additive noise only 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 [5], HOVTS is used to approximate the nonlinear portion of the distortion function by expanding with respect to $\mathbf{n} - \mathbf{x}$ instead of $(\mathbf{x}, \mathbf{n})$. Both approaches work for each feature dimension independently by ignoring correlations among different channels of filterbank. In [16], 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 [6], we extended the above works in the following ways: 1) the nonlinear distortion function for both additive noise and convolutional distortion can be approximated by HOVTS with any order, 2) the required sufficient statistics are derived for estimating model parameters of additive noise and convolutional distortion, and clean speech feature vector, 3) correlations among different channels of filterbank can be considered. So far, we have only published in [7] the formulation for dealing with additive noises, and the corresponding evaluation results on Aurora2 database. In this paper, we present a more general formulation that can deal with both additive noise and convolutional distortion, and report evaluation results on Aurora2, Aurora3, and Aurora4 tasks.

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 report experimental results and finally we conclude the paper in Section 4.

2. 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. 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 $\boldsymbol{\Sigma}_{\mathbf{n}}^c$. Let's further assume that the term $\mathbf{h}^c$ corresponding to convolutional distortion is an unknown deterministic vector. In the recognition stage, the above unknown distortion model parameters can be estimated as follows:

Step 1: Initialization:

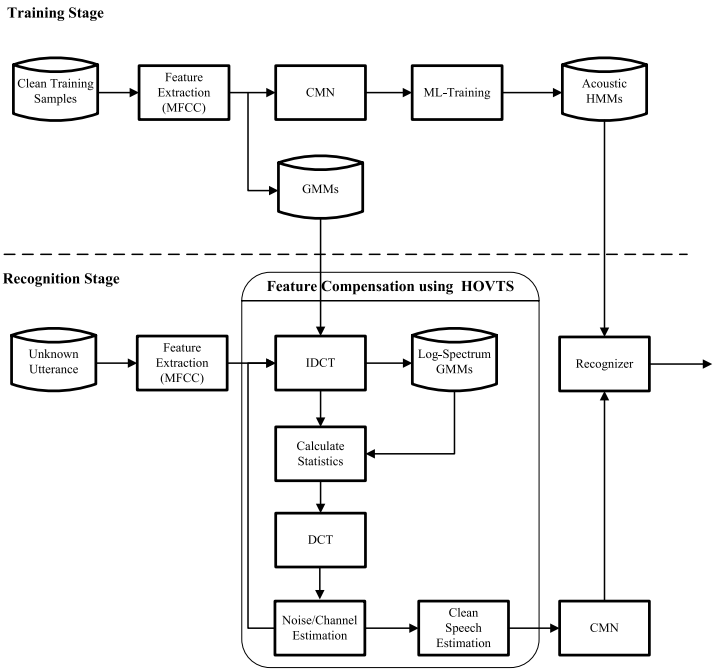

Fig. 1. Flowchart of our feature compensation approach.

We first estimate the initial noise model parameters in cepstral domain by taking the sample mean and covariance of the MFCC features from the first several (10 in our experiments) frames of the unknown utterance, and set $\mathbf{h}^{\mathrm{c}}$ as a zero vector.

Step 2: Define a new random vector, $\mathbf{z}^{c} = \mathbf{x}^{c} + \mathbf{h}^{c}$, whose PDF can be derived as follows:

$$p(\mathbf{z}_t^c) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\mathbf{z}_t^c; \boldsymbol{\mu}_{\mathbf{x},m}^c + \mathbf{h}^c, \boldsymbol{\Sigma}_{\mathbf{x},m}^c).$$

Then transform all parameters from cepstral domain to log-power-spectral domain as follows:

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{z},m}^{\mathrm{l}} &= \mathbf{C}^{+}(\boldsymbol{\mu}_{\mathbf{x},m}^{\mathrm{c}} + \mathbf{h}^{\mathrm{c}}) && (3)\\
\boldsymbol{\Sigma}_{\mathbf{z},m}^{\mathrm{l}} &= \mathbf{C}^{+}\boldsymbol{\Sigma}_{\mathbf{x},m}^{\mathrm{c}}(\mathbf{C}^{+})^{\top} && (4)\\
\boldsymbol{\mu}_{\mathbf{n}}^{\mathrm{l}} &= \mathbf{C}^{+}\boldsymbol{\mu}_{\mathbf{n}}^{\mathrm{c}} && (5)\\
\boldsymbol{\Sigma}_{\mathbf{n}}^{\mathrm{l}} &= \mathbf{C}^{+}\boldsymbol{\Sigma}_{\mathbf{n}}^{\mathrm{c}}(\mathbf{C}^{+})^{\top} && (6)
\end{aligned}$$

where $\mathbf{C}^{+}$ is the Moore-Penrose inverse [10] of the discrete cosine transform (DCT) matrix $\mathbf{C}$, the superscript 'l' and 'c' indicate the log-power-spectral domain and cepstral domain, respectively.

Step 3: In log-power-spectral domain, use HOVTS approximation to calculate the relevant statistics, $\boldsymbol{\mu}_{\mathbf{y},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{y},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{zy},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{ny},m}^{\mathrm{l}}$, which are required for re-estimation of distortion model parameters and estimation of clean speech.

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

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{y},m}^{\mathrm{c}} &= \mathbf{C}\boldsymbol{\mu}_{\mathbf{y},m}^{\mathrm{l}} && (7)\\
\boldsymbol{\Sigma}_{\mathbf{y},m}^{\mathrm{c}} &= \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{y},m}^{\mathrm{l}}(\mathbf{C})^{\top} && (8)\\
\boldsymbol{\Sigma}_{\mathbf{zy},m}^{\mathrm{c}} &= \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{zy},m}^{\mathrm{l}}(\mathbf{C})^{\top} && (9)\\
\boldsymbol{\Sigma}_{\mathbf{ny},m}^{\mathrm{c}} &= \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{ny},m}^{\mathrm{l}}(\mathbf{C})^{\top}\,. && (10)
\end{aligned}$$

Step 5: Use the following updating formulas (extended from e.g., [15, 12]) to re-estimate the distortion model parameters:

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

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

$$\overline{\mathbf{h}} = \left[\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)\boldsymbol{\Sigma}_{\mathbf{x},m}^{-1}\right]^{-1} \left[\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)\boldsymbol{\Sigma}_{\mathbf{x},m}^{-1}(E_{\mathbf{z}}[\mathbf{z}_t|\mathbf{y}_t, m] - \boldsymbol{\mu}_{\mathbf{x},m})\right] \quad (13)$$

where

$$P(m|\mathbf{y}_t) = \frac{\omega_m p_{\mathbf{y}}(\mathbf{y}_t|m)}{\sum_{l=1}^{M}\omega_l p_{\mathbf{y}}(\mathbf{y}_t|l)}\,. \quad (14)$$

In the above equations, we have dropped the cepstral domain indicator "c" in relevant variables for notational convenience. Furthermore, $p_{\mathbf{y}}(\mathbf{y}_t) = \sum_{m=1}^{M}\omega_m p_{\mathbf{y}}(\mathbf{y}_t|m)$, is the PDF of the noisy speech $\mathbf{y}_t$, where the true $p_{\mathbf{y}}(\mathbf{y}_t|m)$ is approximated by a Gaussian PDF, $\mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_{\mathbf{y},m}, \boldsymbol{\Sigma}_{\mathbf{y},m})$, via "moment-matching". $E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t, m]$, $E_{\mathbf{n}}[\mathbf{n}_t\mathbf{n}_t^{\top}|\mathbf{y}_t, m]$ and $E_{\mathbf{z}}[\mathbf{z}_t|\mathbf{y}_t, m]$ are the relevant conditional expectations evaluated as follows:

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

$$E_{\mathbf{n}}[\mathbf{n}_t\mathbf{n}_t^{\top}|\mathbf{y}_t, m] = E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t, m]E_{\mathbf{n}}^{\top}[\mathbf{n}_t|\mathbf{y}_t, m] + \boldsymbol{\Sigma}_{\mathbf{n}} - \boldsymbol{\Sigma}_{\mathbf{ny},m}\boldsymbol{\Sigma}_{\mathbf{y},m}^{-1}\boldsymbol{\Sigma}_{\mathbf{yn},m} \quad (16)$$

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

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

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

$$\hat{\mathbf{x}}_t = E_{\mathbf{x}}[\mathbf{x}_t|\mathbf{y}_t] = \sum_{m=1}^{M} P(m|\mathbf{y}_t) E_{\mathbf{x}}[\mathbf{x}_t|\mathbf{y}_t, m] \quad (18)$$

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:

$$E_{\mathbf{x}}[\mathbf{x}_t|\mathbf{y}_t, m] = E_{\mathbf{z}}[\mathbf{z}_t|\mathbf{y}_t, m] - \mathbf{h}\,. \quad (19)$$

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

To calculate the required statistics, $\boldsymbol{\mu}_{\mathbf{y},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{y},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{zy},m}^{\mathrm{l}}$, $\boldsymbol{\Sigma}_{\mathbf{ny},m}^{\mathrm{l}}$, readers are referred to Section 3 of our previous paper [7]. A small modification, namely replacing x by z in the relevant equations, has to be made though.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

In order to verify the effectiveness of the proposed approach, a series of experiments are performed for the small-vocabulary task of recognition of connected digit strings on Aurora2 and Aurora3 databases, and the large-vocabulary continuous speech recognition (LVCSR) task on Aurora4 database. The Aurora2 and Aurora4 databases contains speech data in the presence of additive noises and linear convolutional distortions, which were introduced synthetically to "clean" speech derived from TIDigits and WSJ databases, respectively. The Aurora3 database contains utterances of digit strings recorded in real automobile environments for Danish, Finnish, German and Spanish, respectively. A full description of the above databases and the corresponding test frameworks are given in [8, 1, 2, 3, 4, 9, 14].

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 spectrums. The mixture number of clean-speech GMM for feature compensation is 256. For Aurora2 and Aurora3 tasks, each digit is modeled by a whole-word left-to-right CDHMM, which consists of 16 emitting states, each having 3 Gaussian mixture components. For Aurora4 task, triphones are used as basic speech units. Each triphone is modeled by a CDHMM with 3 emitting states, each having 8 Gaussian mixture components. There are in total 2800 tied states based on decision trees. A bigram language model (LM) for a 5k-word vocabulary is used in recognition.

For experiments on Aurora2 and Aurora4 databses, "clean-training" is used, where 8kHz data is used for Aurora4. For Aurora3 experiments, we focus on high-mismatch (HM) "training-testing" condition, where training data includes utterances recorded by close-talking (CT) microphone, which can be considered as "clean", while testing data is recorded by hands-free (HF) microphone. For re-estimation of distortion model parameters, 4 EM iterations are used. Our baseline systems used CMN for feature compensation. In all the experiments, tools in HTK [17] are used for training and testing.

Experiments are designed to compare the following three HOVTS-based methods:

- **VTS(N)**: only consider additive noise (i.e., the method in [7]);
- **VTS(N,H)**: consider both additive noise and convolutional distortion as described in Section 2;
- **CMN+VTS(N,H)**: CMN is applied first before using "VTS(N,H)" for additional feature compensation. In this case, the clean-speech GMM is also trained using the CMN-processed MFCCs.

In the following subsection, we report the experimental results.

3.2. 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 on Aurora2 database. 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. Higher the order in VTS approximation, better the performance, especially in VTS(N) case. Second, 3rd-order VTS(N) achieves the best overall performance. But for the channel mismatch case (i.e., Set C), VTS(N,H) is better, which indicates that channel re-estimation is useful. Third, CMN+VTS(N,H) performs worse than VTS(N,H).

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		66.36	71.43	67.20	68.55
VTS(N)	1st-order	86.21	85.24	82.65	85.11
	2nd-order	87.18	86.61	84.90	86.49
	3rd-order	87.65	87.01	85.39	86.94
VTS(N,H)	1st-order	86.37	84.72	84.17	85.27
	2nd-order	86.97	85.70	85.14	86.09
	3rd-order	87.32	86.13	85.62	86.50
CMN + VTS(N,H)	1st-order	85.22	84.60	84.07	84.74
	2nd-order	84.95	85.15	84.52	84.94
	3rd-order	85.41	85.49	84.77	85.32

Table 2. Performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using HOVTS-based feature compensation in the high-mismatch (HM) condition on Aurora3 database.

Methods		German	Danish	Finnish	Spanish
Baseline		83.77	54.78	77.07	80.96
VTS(N)	1st	88.85	46.24	82.83	69.95
	2nd	89.73	57.01	84.45	77.95
	3rd	90.06	61.01	84.06	78.41
VTS(N,H)	1st	89.87	56.86	83.50	77.83
	2nd	90.01	63.40	84.81	79.46
	3rd	90.33	66.69	84.66	79.46
CMN + VTS(N,H)	1st	89.59	74.18	84.81	84.39
	2nd	89.96	72.14	85.16	84.27
	3rd	90.43	73.86	85.69	85.05

Table 2 summarizes a performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using HOVTS-based feature compensation in the high-mismatch (HM) condition on Aurora3 database. We made the following observations:

- For first-order VTS, channel re-estimation is important, especially on Danish and Spanish databases;
- CMN+VTS(N,H) outperforms VTS(N) and VTS(N,H);
- On Spanish database, the performance of VTS(N) and VTS(N,H) is even worse than the baseline system. The assumed distortion model seems not good enough to characterize the true distortions in this case. However, by applying CMN first, the HOVTS-based feature compensation works.

Tables 3 and 4 summarize a performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using HOVTS-based feature compensation for two different microphones on Aurora4 database. The Sennheiser microphone in Table 3 is also used for recording "clean-training" data. So the results in Table 4 are used for demonstrating the effects of both additive noises and channel mismatch. It is observed that HOVTS-based feature compensation helps, although the gain is not as big as that achieved for Auroa2 and Aurora3 tasks. Again CMN+VTS(N,H) achieves the best performance.

Table 3. Performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using HOVTS-based feature compensation for the Sennheiser microphone on Aurora4 database.

Methods		Car	Babble	Restaurant	Street	Airport	Train	Overall
Baseline		76.09	52.87	51.30	49.36	54.77	48.05	55.41
VTS(N)	1st-order	79.02	58.73	53.63	55.11	59.93	57.43	60.64
	2nd-order	75.45	60.90	56.12	57.20	61.07	60.47	61.86
	3rd-order	75.64	61.14	57.05	58.08	61.80	60.60	62.38
VTS(N,H)	1st-order	79.25	59.69	53.76	55.39	59.39	58.40	60.98
	2nd-order	76.41	59.65	55.31	56.53	59.80	59.41	61.18
	3rd-order	76.44	60.81	56.29	57.67	61.01	59.91	62.02
CMN + VTS(N,H)	1st-order	78.61	61.26	55.35	57.44	60.38	61.16	62.36
	2nd-order	76.24	62.75	56.77	58.27	60.94	61.07	62.67
	3rd-order	76.72	62.99	57.24	59.22	61.40	61.83	63.23

Table 4. Performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using HOVTS-based feature compensation for the second microphone on Aurora4 database.

Methods		Car	Babble	Restaurant	Street	Airport	Train	Overall
Baseline		65.20	44.22	40.48	36.07	44.80	37.96	44.79
VTS(N)	1st-order	73.42	51.37	45.66	46.24	53.63	51.84	53.69
	2nd-order	74.48	55.61	48.51	51.02	56.42	54.81	56.81
	3rd-order	74.03	57.35	49.95	50.64	57.24	56.06	57.55
VTS(N,H)	1st-order	77.43	56.59	47.45	51.28	55.99	57.61	57.73
	2nd-order	76.57	57.58	50.85	52.05	55.95	56.73	58.29
	3rd-order	76.18	58.38	50.46	52.08	57.28	56.72	58.52
CMN + VTS(N,H)	1st-order	77.17	56.88	50.42	53.67	57.71	57.09	58.82
	2nd-order	76.82	58.81	52.38	54.25	58.36	57.84	59.74
	3rd-order	76.80	58.64	52.46	54.27	58.55	57.65	59.73

4. CONCLUSION

From the above experimental results, mixed observations can be made for different tasks on different databases. Although performance improvement is achieved in many cases by using a higher order VTS-based feature compensation compared with the first-order VTS, it also requires more computations. The most useful finding from this study is that the CMN+VTS(N,H) approach works for real data on Aurora3 task. We therefore recommend our readers to try out this approach as well in their applications.

5. REFERENCES

[1] Aurora document AU/217/99, "Availability of Finnish SpeechDat-Car database for ETSI STQ WI008 front-end standardisation," Nokia, Nov. 1999.

[2] Aurora document AU/271/00, "Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: description and baseline results," UPC, Nov. 2000.

[3] Aurora document AU/273/00, "Description and baseline results for the subset of the SpeechDat-Car German database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation," Texas Instruments, Dec. 2001.

[4] Aurora document AU/378/01, "Danish SpeechDat-Car digits database for ETSI STQ-Aurora advanced DSR," Aalborg University, Jan. 2001.

[5] 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.

[6] J. Du and Q. Huo, "Feature compensation using high-order vector Taylor series for noisy speech recognition," Technical Memo, MSRA, January 2008.

[7] J. Du and Q. Huo, "A feature compensation approach using high-order vector Taylor series approximation of an explicit distortion model for noisy speech recognition," to appear in *Proc. Interspeech*, 2008.

[8] 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.

[9] H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, version 2.0," ETSI STQ-Aurora DSR Working Group, Nov. 2002.

[10] 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.

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

[12] 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.

[13] 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.

[14] N. Parihar and J. Picone, "DSR front end LVCSR evaluation," AU/384/02, Aurora Working Group, ETSI, Dec. 2002.

[15] 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.

[16] V. Stouten, *Robust automatic speech recognition in time-varying environemnts*, Ph.D. thesis, Katholieke Universiteit Leuven, 2006.

[17] S. Young *et al.*, The HTK Book (for HTK v3.4), 2006.