A VTS-BASED FEATURE COMPENSATION APPROACH TO NOISY SPEECH RECOGNITION USING MIXTURE MODELS OF DISTORTION

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

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

ABSTRACT

Recently, we proposed an approach to irrelevant variability normalization (IVN) based joint training of a reference Gaussian mixture model (GMM) for feature compensation and hidden Markov models (HMMs) for acoustic modeling by using a vector Taylor series (VTS) based feature compensation technique, where single-component densities are used to model additive noise and convolutional distortion respectively. In this paper, mixtures of densities are used to enhance the distortion model. New formulations for maximum likelihood (ML) estimation of distortion model parameters, and minimum mean squared error (MMSE) estimation of clean speech are derived and presented. A comparative study is conducted under three "training-testing" conditions on Aurora3 database. Experimental results confirm that the proposed mixture models of distortion can achieve significant performance gain compared with the traditional distortion modeling.

Index Terms— feature compensation, vector Taylor series, mixture model of distortion, irrelevant variability normalization

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 dealing with the above problem is the so-called feature compensation approach using *explicit* model of environmental distortions (e.g., [1]), which is also the topic of this paper.

For our approach, it is assumed that in time domain, "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. In log-power-spectral domain, the distortion model can be expressed *approximately* (e.g., [1]) as

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

*This work was done when Jun Du worked at the Speech Group of Microsoft Research Asia, Beijing, China.

where $\mathbf{y}^{\text{l}}$, $\mathbf{x}^{\text{l}}$, $\mathbf{h}^{\text{l}}$ and $\mathbf{n}^{\text{l}}$ are log power-spectra of noisy speech, clean speech, convolutional term and noise, respectively. In MFCC domain, the distortion model becomes

$$\mathbf{y}^{\text{c}} = \mathbf{C}\log[\exp(\mathbf{C}^{+}(\mathbf{x}^{\text{c}} + \mathbf{h}^{\text{c}})) + \exp(\mathbf{C}^{+}\mathbf{n}^{\text{c}})] \tag{3}$$

where $\mathbf{C}$ is a $D^{\text{c}} \times D^{\text{l}}$ truncated discrete cosine transform (DCT) matrix, $\mathbf{C}^{+}$ denotes the Moore-Penrose inverse of $\mathbf{C}$ (refer to [2] for details), D^{c} is the dimension of MFCC feature vector, and D^{l} is the number of channels of the Mel-frequency filterbank used in MFCC feature extraction. In most of current ASR systems, $D^{\text{c}} < D^{\text{l}}$. The log and exp functions in the above equations operate element-by-element on the corresponding vectors. 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 first-order vector Taylor series (VTS) approximation, has been tried in the past (e.g., [3, 4]). The related works to VTS-based feature compensation can be divided into several categories. The first category is on the more precise expression of distortion model in Eq. (2). An example is given in [5], where the phase relationship between clean speech and additive noise is incorporated into the distortion model. The second category is on the more accurate approximation of the nonlinear distortion model. In [6], a linear function is found to approximate the high-order Taylor series expansion of the above nonlinear distortion model by minimizing the mean-squared error. In [7], the nonlinear distortion model is approximated by a second-order VTS. More recently, we proposed a high-order VTS based formulation for maximum likelihood (ML) estimation of both additive noise and convolutional distortion, and minimum mean squared error (MMSE) estimation of clean speech in [8]. The third category is on improving the recognition accuracy in non-stationary environments. In [5, 7], sequential noise estimation is performed to deal with non-stationary noise. The last category is on the extension from traditional VTS-based feature compensation under the clean-training condition to real scenarios, where noisy speech can also be included in the training data. In [9], noise adaptive training (NAT) (e.g., [10]) was used to train a front-end GMM. Meanwhile, we proposed to use irrelevant variability normalization (IVN) based joint training of a reference Gaussian mixture model (GMM) for feature compensation and HMMs for acoustic modeling in [11], which outperforms the method using IVN-based training [2] (similar to NAT) of the GMM.

The main contribution of this paper is to adopt a mixture model for modeling both additive noise and convolutional distortion, which is combined with our recently proposed IVN-based joint training of GMM and HMMs using VTS-based feature compensation, to im-

prove the recognition accuracy in non-stationary environments. The corresponding formulations for ML estimation of distortion model parameters, and MMSE estimation of clean speech are derived and presented here. Our work is related to a recent work in [12], where a similar idea of using noise mixture model is proposed. But the method in [12] significantly differs from ours in this paper. First, the operation domain of VTS is Log-Mel-Filter-Bank (LMFB) in [12] while we operate on MFCCs which are the final output features fed to the recognizer. Second, mixture model is only used for additive noise and the estimation of noise mixture model and bias vector (i.e., convolutional distortion in this paper) is in an alternate manner of switching between different auxiliary functions by using MMSE estimator of clean speech and noise in [12], while we use mixture models for both additive noise and convolutional distortion and closed-form formulations can be derived by jointly optimizing all the parameters of the distortion model using a unique auxiliary function for ML estimation. Furthermore, our formulations are generalized to VTS with any order. Third, for the noise suppression, a Mel-scaled Wiener filter is exploited in [12] while we use MMSE estimation of clean speech. Finally, the method is verified under clean-training condition where VTS-based feature compensation is only performed on the testing set with synthesized noisy speech in [12], while we use IVN-based joint training to extend VTS-based feature compensation to any "training-testing" condition and verify our approach on noisy speech from real environments.

The rest of the paper is organized as follows. In Section 2, we introduce VTS-based feature compensation using mixture models of distortion. In Section 3, we review the procedure for IVN-based joint training of GMM and HMMs using VTS-based feature compensation. In Section 4, we report experimental results. Finally, we conclude the paper in Section 5.

2. VTS-BASED FEATURE COMPENSATION USING MIXTURE MODELS OF DISTORTION

In [8], the clean speech is modeled by a GMM as follows:

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

For each utterance, we assume that the additive noise $\mathbf{n}^c$ follows a Gaussian PDF (probability density function) while the convolutional distortion $\mathbf{h}^c$ has a PDF of the Kronecker delta function. In this work, to enhance the modeling power for distortions, mixture models are employed to model both additive noise and convolutional distortion as follows:

$$p(\mathbf{n}_t^c) = \sum_{l=1}^{L} \omega_{\mathbf{n},l} \mathcal{N}(\mathbf{n}_t^c; \boldsymbol{\mu}_{\mathbf{n},l}^c, \boldsymbol{\Sigma}_{\mathbf{n},l}^c) \quad (4)$$

$$p(\mathbf{h}_t^c) = \sum_{k=1}^{K} \omega_{\mathbf{h},k} \delta(\mathbf{h}_t^c - \mathbf{h}_{\text{const},k}^c) \quad (5)$$

where GMM and mixture of Kronecker delta functions are used for modeling additive noise and convolutional distortion, respectively. In our implementation, the mixture number of additive noise L is set equal to the mixture number of convolutional distortion K as we assume that each pair of mixture component can roughly model a stationary segment of an utterance. Also we should define a new random vector, $\mathbf{z}^c = \mathbf{x}^c + \mathbf{h}^c$, whose PDF can be derived as follows:

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

The above unknown distortion model parameters can be estimated as follows:

Step 1: *Initialization*

For each utterance, first we determine the mixture number by setting $L = K = \left\lceil \frac{T}{T_{Seg}} \right\rceil$, where T_{Seg} and T are the length of a relatively stationary segment and the current utterance, respectively. Then we use the procedure in [8] to estimate a global set of distortion parameters to initialize each pair of mixture components, followed by the parameter re-estimation in [8] using frames of each segment corresponding to each mixture component separately. All mixture weights are set to equal.

Step 2: *Computation of required statistics*

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

$$\boldsymbol{\mu}_{\mathbf{z},mk}^{\text{l}} = \mathbf{C}^+(\boldsymbol{\mu}_{\mathbf{x},m}^{\text{c}} + \mathbf{h}_{\text{const},k}^{\text{c}}) \quad (6)$$

$$\boldsymbol{\Sigma}_{\mathbf{z},m}^{\text{l}} = \mathbf{C}^+ \boldsymbol{\Sigma}_{\mathbf{x},m}^{\text{c}} (\mathbf{C}^+)^\top \quad (7)$$

$$\boldsymbol{\mu}_{\mathbf{n},l}^{\text{l}} = \mathbf{C}^+ \boldsymbol{\mu}_{\mathbf{n},l}^{\text{c}} \quad (8)$$

$$\boldsymbol{\Sigma}_{\mathbf{n},l}^{\text{l}} = \mathbf{C}^+ \boldsymbol{\Sigma}_{\mathbf{n},l}^{\text{c}} (\mathbf{C}^+)^\top \quad (9)$$

where the superscripts 'l' and 'c' indicate the log-power-spectral domain and cepstral domain, respectively. Then in log-power-spectral domain, use high-order VTS approximation [8] to calculate the relevant statistics , $\boldsymbol{\mu}_{\mathbf{y},mkl}^{\text{l}}$, $\boldsymbol{\Sigma}_{\mathbf{y},mkl}^{\text{l}}$, $\boldsymbol{\Sigma}_{\mathbf{zy},mkl}^{\text{l}}$, $\boldsymbol{\Sigma}_{\mathbf{ny},mkl}^{\text{l}}$, which are required for re-estimation of distortion model parameters and estimation of clean speech. Finally, transform the statistics back to cepstral domain as follows:

$$\boldsymbol{\mu}_{\mathbf{y},mkl}^{\text{c}} = \mathbf{C}\boldsymbol{\mu}_{\mathbf{y},mkl}^{\text{l}} \quad (10)$$

$$\boldsymbol{\Sigma}_{\mathbf{y},mkl}^{\text{c}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{y},mkl}^{\text{l}} (\mathbf{C})^\top \quad (11)$$

$$\boldsymbol{\Sigma}_{\mathbf{zy},mkl}^{\text{c}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{zy},mkl}^{\text{l}} (\mathbf{C})^\top \quad (12)$$

$$\boldsymbol{\Sigma}_{\mathbf{ny},mkl}^{\text{c}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{ny},mkl}^{\text{l}} (\mathbf{C})^\top \ . \quad (13)$$

Step 3: *Joint re-estimation of distortion model parameters*

Use Eq. (14) to Eq. (18) to re-estimate the distortion model parameters. Note that the cepstral domain indicator "c" in relevant variables has been dropped for notational convenience. The detailed derivations for joint re-estimation will be reported elsewhere, which can be extended from those in [13, 8]. Several items used in Eq. (14) to Eq. (18) are evaluated in Eq. (19) to Eq. (22), where the statistics $\boldsymbol{\mu}_{\mathbf{y},mkl}$, $\boldsymbol{\Sigma}_{\mathbf{y},mkl}$, $\boldsymbol{\Sigma}_{\mathbf{zy},mkl}$, $\boldsymbol{\Sigma}_{\mathbf{ny},mkl}$ are calculated in Step 2.

Step 4: *Repeat Step 2 and Step 3 N_{VTS} 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

$$\begin{aligned} \hat{\mathbf{x}}_t &= E_{\mathbf{x}}[\mathbf{x}_t|\mathbf{y}_t] \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{l=1}^{L} P(m,k,l|\mathbf{y}_t)(E_{\mathbf{z}}[\mathbf{z}_t|\mathbf{y}_t,m,k,l] - \mathbf{h}_{\text{const},k}) \ . \end{aligned} \quad (23)$$

$$\overline{\omega}_{\mathbf{n},l} = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{k=1}^{K} P(m,k,l|\mathbf{y}_t) \tag{14}$$

$$\overline{\boldsymbol{\mu}}_{\mathbf{n},l} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{k=1}^{K} P(m,k,l|\mathbf{y}_t) E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t,m,k,l]}{\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{k=1}^{K} P(m,k,l|\mathbf{y}_t)} \tag{15}$$

$$\overline{\boldsymbol{\Sigma}}_{\mathbf{n},l} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{k=1}^{K} P(m,k,l|\mathbf{y}_t) E_{\mathbf{n}}[\mathbf{n}_t\mathbf{n}_t^{\top}|\mathbf{y}_t,m,k,l]}{\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{k=1}^{K} P(m,k,l|\mathbf{y}_t)} - \overline{\boldsymbol{\mu}}_{\mathbf{n},l}\overline{\boldsymbol{\mu}}_{\mathbf{n},l}^{\top} \tag{16}$$

$$\overline{\omega}_{\mathbf{h},k} = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{l=1}^{L} P(m,k,l|\mathbf{y}_t) \tag{17}$$

$$\overline{\mathbf{h}}_{\text{const},k} = \left[\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{l=1}^{L} P(m,k,l|\mathbf{y}_t)\boldsymbol{\Sigma}_{\mathbf{x},m}^{-1}\right]^{-1}\left[\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{l=1}^{L} P(m,k,l|\mathbf{y}_t)\boldsymbol{\Sigma}_{\mathbf{x},m}^{-1}(E_{\mathbf{z}}[\mathbf{z}_t|\mathbf{y}_t,m,k,l] - \boldsymbol{\mu}_{\mathbf{x},m})\right] \tag{18}$$

$$P(m,k,l|\mathbf{y}_t) = \frac{\omega_{\mathbf{x},m}\omega_{\mathbf{h},k}\omega_{\mathbf{n},l}\mathcal{N}(\mathbf{y}_t;\boldsymbol{\mu}_{\mathbf{y},mkl},\boldsymbol{\Sigma}_{\mathbf{y},mkl})}{\sum_{l=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L}\omega_{\mathbf{x},m}\omega_{\mathbf{h},k}\omega_{\mathbf{n},l}\mathcal{N}(\mathbf{y}_t;\boldsymbol{\mu}_{\mathbf{y},mkl},\boldsymbol{\Sigma}_{\mathbf{y},mkl})} \tag{19}$$

$$E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t,m,k,l] = \boldsymbol{\mu}_{\mathbf{n},l} + \boldsymbol{\Sigma}_{\mathbf{ny},mkl}\boldsymbol{\Sigma}_{\mathbf{y},mkl}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},mkl}) \tag{20}$$

$$E_{\mathbf{n}}[\mathbf{n}_t\mathbf{n}_t^{\top}|\mathbf{y}_t,m,k,l] = E_{\mathbf{n}}[\mathbf{n}_t|\mathbf{y}_t,m,k,l]E_{\mathbf{n}}^{\top}[\mathbf{n}_t|\mathbf{y}_t,m,k,l] + \boldsymbol{\Sigma}_{\mathbf{n},l} - \boldsymbol{\Sigma}_{\mathbf{ny},mkl}\boldsymbol{\Sigma}_{\mathbf{y},mkl}^{-1}\boldsymbol{\Sigma}_{\mathbf{yn},mkl} \tag{21}$$

$$E_{\mathbf{z}}[\mathbf{z}_t|\mathbf{y}_t,m,k,l] = (\boldsymbol{\mu}_{\mathbf{x},m} + \mathbf{h}_{\text{const},k}) + \boldsymbol{\Sigma}_{\mathbf{zy},mkl}\boldsymbol{\Sigma}_{\mathbf{y},mkl}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},mkl}) \tag{22}$$

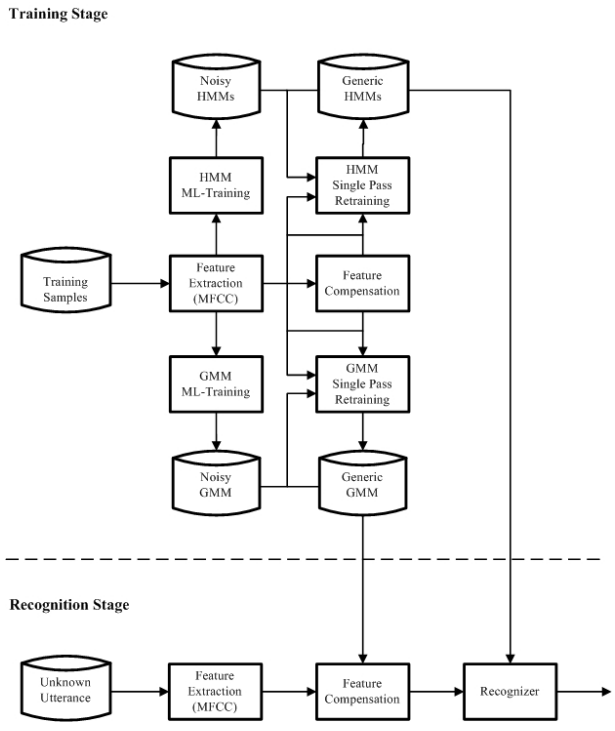

Fig. 1. Flowchart of IVN training using VTS-based feature compensation.

3. IVN-BASED JOINT TRAINING OF GMM AND HMMS

In the traditional framework of VTS-based feature compensation, both HMMs for recognition and reference GMM for feature compensation are trained on clean speech data. In real scenarios, the training data may include noisy speech data. In [11], we propose a procedure to perform IVN-based joint training of GMM and HMMs using VTS-based feature compensation, which is illustrated in Fig. 1. In the training stage, the procedure is as follows:

Step 1: *Initialization*
First, the reference GMM for feature compensation and HMMs for recognition are ML-trained from training data using MFCC features with cepstral mean normalization (CMN).

Step 2: *VTS-based feature compensation*
Given the GMM, VTS-based feature compensation is applied to each training utterance.

Step 3: *Joint training of GMM and HMMs*
Based on the compensated features of training set, single pass retraining (SPR) [14] is performed to generate the generic GMM and HMMs by using the last updated GMM and HMMs with the corresponding feature set. The SPR works as follows: given one set of well-trained models, a new set matching a different training data parameterization can be generated in a single re-estimation pass, which is done by computing the forward and backward probabilities using the original models together with the original training data and then switching to the new training data to compute the parameter estimation for the new set of models.

Step 4: *Repeat Step 2 and Step 3 N_{IVN} times*.

In the recognition stage, after feature extraction for an unknown utterance, we perform VTS-based feature compensation using generic GMM and then do recognition using generic HMMs.

Table 1. Performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using VTS-based feature compensation under three "training-testing" conditions on Aurora3 database.

Methods		German	Danish	Finnish	Spanish
Baseline	HM	83.77	54.78	77.07	80.99
	MM	82.43	75.42	84.06	89.39
	WM	92.49	90.84	93.09	93.57
VTS	HM	91.77	77.39	90.46	87.46
IVN-VTS	HM	92.09	79.98	91.55	88.90
	MM	89.24	78.53	87.48	91.44
	WM	94.93	92.91	95.64	95.57
IVN-MMD-VTS	HM	92.74	80.64	92.61	91.28
	MM	89.70	80.08	88.30	91.97
	WM	95.13	93.05	96.23	95.78

In the above procedure, the IVN concept is implemented by SPR using VTS-based feature compensation. Actually, there are other two alternatives which can also achieve this goal. One method is to use the compensated features to retrain GMM from scratch and then use the new GMM to compensate features again in an iterative way. Finally a generic GMM can be generated. The other method is to use a similar procedure as in [2] to generate a generic GMM. For those two methods, the generic HMMs can be trained from scratch using compensated features based on generic GMM. As a comparison, our SPR-based IVN training has two advantages: 1) GMM and HMMs are jointly trained in each iteration, 2) both GMM and HMMs are progressively updated, which brings stable improvements of recognition performance. Our experimental results also confirm that SPR-based IVN training can achieve better recognition performance, which is recommended as a practical solution. In our previous work, the effectiveness of IVN-based joint training is verified only under the well-matched "training-testing" condition. In this work, we will give a comprehensive performance comparison under different "training-testing" conditions, where some interesting observations are made.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

In order to verify the effectiveness of the proposed approach on real-world ASR, Aurora3 databases are used, which contain utterances of digit strings recorded in real automobile environments for German, Danish, Finnish and Spanish, respectively. A full description of the above databases and the corresponding test frameworks are given in [15, 16, 17, 18].

In our ASR systems, each feature vector 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. Three "training-testing" conditions are designed for Aurora3. The first one is high-mismatch (HM) 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. The second one is well-matched (WM) condition, where both training and testing data are recorded by CT and HF microphones. The last one is mid-mismatch (MM) condition which can be considered as the tradeoff between WM condition and HM condition. The relevant control parameters are set as $M = 256$, $T_{Seg} = 60$, $N_{VTS} = 4$, $N_{IVN} = 4$. Other control parameters related to our previous work on VTS-based feature compensation can be found in [8, 11]. Our baseline system uses cepstral mean normalization (CMN) for feature compensation. In all the experiments, tools in HTK [14] are used for training and testing.

4.2. Experimental Results

Table 1 summarizes a performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using VTS-based feature compensation under three "training-testing" conditions (HM, MM, WM) on Aurora3 database. VTS refers to the practical solution of feature compensation recommended in [8] with two additional improvements, i.e., applying higher order information of VTS approximation only to the noisy speech mean parameters and acoustic context expansion in [11] under the clean-training condition. IVN-VTS represents the system where IVN training with VTS-based feature compensation is used. IVN-MMD-VTS denotes the system using mixture models of distortion for VTS-based feature compensation combined with IVN training. From those results, several observations can be made as follows:

- All robust ASR systems using VTS-based feature compensation outperform the baseline system under all "training-testing" conditions for four languages;
- Under the HM condition, although we treat it as "clean-training" condition where VTS-based feature compensation can only be applied to the testing set, IVN-VTS system can still achieve consistent improvement of recognition accuracy over the VTS system;
- IVN-MMD-VTS system yields consistent and significant gain of recognition accuracy compared with IVN-VTS system for all the testing cases.

5. CONCLUSION

In this paper, we propose to use mixture models for modeling both additive noise and convolutional distortion to improve the recognition accuracy in non-stationary environments. Combined with IVN-based joint training of a reference GMM for feature compensation and HMMs for acoustic modeling using VTS-based feature compensation, significant performance gain can be achieved under all the "training-testing" conditions on Aurora3 task.

6. REFERENCES

[1] A. Acero, *Acoustic and Environment Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1993.

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

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

[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] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 2, pp. 133–143, 2004.

[6] N. S. Kim, "Statistical linear approximation for environment compensation," *IEEE Signal Processing Letters*, vol. 5, no. 1, pp. 8–10, 1998.

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

[8] 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," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2285–2293, 2011.

[9] J. Li, M. L. Seltzer, and Y. Gong, "Improvements to VTS feature enhancement," in *ICASSP*, 2012, pp. 4677–4680.

[10] O. Kalinli, M. L. Seltzer, and A. Acero, "Noise adaptive training using a vector Taylor series approach for robust automatic speech recognition," in *ICASSP*, 2009, pp. 3825–3828.

[11] J. Du and Q. Huo, "IVN-based joint training of GMM and HMMs using an improved VTS-based feature compensation for noisy speech recognition," in *Interspeech*, 2012.

[12] M. Fujimoto, S. Watanabe, and T. Nakatani, "Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation," in *ICASSP*, 2012, pp. 4713–4716.

[13] 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.

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

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

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

[17] 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, 2001.

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