ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

# IVN-Based Joint Training Of GMM And HMMs Using An Improved VTS-Based Feature Compensation For Noisy Speech Recognition

*Jun Du, Qiang Huo*

Microsoft Research Asia, Beijing, P. R. China
{jundu, qianghuo}@microsoft.com

## Abstract

In our previous work, we proposed a feature compensation approach using high-order vector Taylor series approximation for noisy speech recognition. In this paper, first we improve the feature compensation in both efficiency and accuracy by boosted mixture learning of GMM, applying higher order information of VTS approximation only to the noisy speech mean parameters, acoustic context expansion, and modeling the convolutional distortion as a single Gaussian. Then we design a procedure to perform irrelevant variability normalization based joint training of GMM and HMM using the improved VTS-based feature compensation. The effectiveness of our proposed approach is confirmed by experiments on Aurora3 tasks.

***Index Terms***— irrelevant variability normalization, feature compensation, vector Taylor series.

## 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 $t^{\text{th}}$ 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

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

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 [9] for details), $D^{\text{c}}$ is the dimension of MFCC feature vector, and $D^{\text{l}}$ 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., [11, 10]). In our previous work [6], we extend feature compensation from traditional first-order VTS to high-order VTS with any order and give the corresponding re-estimation formulations of parameters for both noise and convolutional distortion. In this paper, we first improve the feature compensation in both efficiency and accuracy as follows: 1) the reference (clean) GMM is built by boosted mixture learning (BML) [7], 2) higher order information of VTS approximation is only applied to the noisy speech mean parameters, which can be very efficient and lead to more stable improvement of recognition accuracy, 3) the posterior probability of mixture component given each frame is calculated by a weighted average among neighboring frames to leverage acoustic context expansion in MMSE estimation of clean speech, 4) the convolutional distortion is modeled by a single Gaussian instead of Kronecker delta function (i.e., a constant). Then we design a procedure to perform irrelevant variability normalization (IVN) based joint training of GMM and HMM using VTS-based feature compensation.

The rest of the paper is organized as follows. In Section 2, we introduce an improved approach to VTS-based feature compensation. In Section 3, we present the detailed procedure for IVN-based joint training of GMM and HMM using VTS-based feature compensation. In Section 4, we report experimental results. Finally, we conclude the paper in Section 5.

## 2. Improved VTS-based Feature Compensation

### 2.1. Boosted Mixture Learning Of Reference GMM

First, we perform boosted mixture learning (BML) of our reference (clean) GMM. This is motivated by a recent work [7] on BML of CDHMMs based on maximum likelihood for speech recognition. BML is an incremental method to learn a mixture model, where in each step one new mixture component is estimated according to the functional gradient of an objective function to ensure that it is added along the direction that maximizes the objective function. In [7], BML achieves significant improvements of recognition performance over the conventional ML training procedure, especially for small model sizes. So for our VTS-based feature compensation, it's natural to train

the reference GMM by using BML, which is verified to achieve better recognition performance. One important issue to make BML effective is the initialization of sample weights in each step of increasing mixture component. In this work, we use one of the methods in [7], namely sampling boosting, to initialize the sample weights.

### 2.2. Use Of Higher Order Information

Second, higher order information of VTS approximation is only applied to calculation of noisy speech mean parameters. In [6], high-order VTS approximation of the nonlinear distortion function is applied to the calculation of all required statistics in log-power-spectral domain, including noisy speech mean and variance parameters, and other covariance parameters. But inconsistent improvements of recognition performance are observed on different Aurora3 tasks. Our new experiments show that if high-order VTS approximation is only applied to noisy speech mean parameters, consistent improvements of recognition performance can be achieved, yet its computational complexity is much lower than that of the original high-order VTS so that the additional computation cost can be ignored compared with full operations of first-order VTS.

### 2.3. Acoustic Context Expansion

Third, we use acoustic context expansion in clean speech estimation to further improve the accuracy. Acoustic context expansion has been effective in several feature extraction/transformation methods, such as TANDEM [8] and fMPE [12], where in addition to the current frame, the information from several neighboring frames in the left and right context is also used. In our VTS-based feature compensation, given the noisy speech feature vector of the $t^{\text{th}}$ frame $\mathbf{y}_t$ and the estimated distortion model parameters, the minimum mean-squared error (MMSE) estimation of clean speech feature vector $\mathbf{x}_t$ in cepstral domain is calculated as [6]

$$\hat{\mathbf{x}}_t = E_{\mathbf{x}}\left[\mathbf{x}_t|\mathbf{y}_t\right] = \sum_{m=1}^{M} P(m|\mathbf{y}_t) E_{\mathbf{x}}\left[\mathbf{x}_t|\mathbf{y}_t, m\right] \ . \quad (4)$$

To leverage acoustic context expansion, we calculate the new posterior probability by a weighted average among neighboring frames as follows:

$$[P(m|\mathbf{y}_t)]_{\text{new}} = \frac{\sum_{\tau=-\Delta}^{\Delta}(\Delta+1-|\tau|)P(m|\mathbf{y}_{t+\tau})}{\sum_{\tau=-\Delta}^{\Delta}(\Delta+1-|\tau|)} \quad (5)$$

where $\Delta$ is the size for context expansion.

### 2.4. Gaussian Assumption For Convolutional Distortion

In [6], the convolutional distortion $\mathbf{h}$ has a probability density function (pdf) of the Kronecker delta function $\delta(\mathbf{h}-\mathbf{h}_{\text{const}})$, where $\mathbf{h}_{\text{const}}$ is an unknown deterministic vector. In this work, to model the variation of convolutional distortion in an utterance, we assume $\mathbf{h}$ follows a Gaussian pdf with a mean vector $\boldsymbol{\mu}_{\mathbf{h}}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_{\mathbf{h}}$. To derive a closed-form solution for ML re-estimation of both noise and convolutional distortion under this assumption, first a likelihood function is defined as follows:

$$\mathcal{L}(\mathbf{Y}|\boldsymbol{\Lambda}) = \sum_{\mathbf{M_x}}\sum_{\mathbf{M_n}}\sum_{\mathbf{M_h}} p(\mathbf{Y}, \mathbf{M_x}, \mathbf{M_n}, \mathbf{M_h}|\boldsymbol{\Lambda_x}, \boldsymbol{\Lambda_n}, \boldsymbol{\Lambda_h}) \quad (6)$$

where $\boldsymbol{\Lambda_x}$, $\boldsymbol{\Lambda_n}$, and $\boldsymbol{\Lambda_h}$ are model parameter sets for $\mathbf{x}$, $\mathbf{n}$ and $\mathbf{h}$, respectively. $\mathbf{Y}$ is the sequence of the noisy observation vectors in the current utterance. $\mathbf{M_x}$, $\mathbf{M_n}$, and $\mathbf{M_h}$ are the sequences of Gaussian component indices for $\mathbf{x}$, $\mathbf{n}$, and $\mathbf{h}$, respectively. Then we adopt an iterative EM algorithm to solve the problem. The M-Step of the EM algorithm is to maximize the following auxiliary function:

$$\mathcal{Q}(\bar{\boldsymbol{\Lambda}}|\boldsymbol{\Lambda}) = E[\log p(\mathbf{X}, \mathbf{N}, \mathbf{H}, \mathbf{M_x}, \mathbf{M_n}, \mathbf{M_h}|\bar{\boldsymbol{\Lambda}})|\mathbf{X}, \mathbf{N}, \mathbf{H}, \boldsymbol{\Lambda}] \quad (7)$$

where $\boldsymbol{\Lambda}$ and $\bar{\boldsymbol{\Lambda}}$ are the sets of old and new model parameters, respectively. The other details of derivation are similar to what is described in the appendix of [6]. Finally the formulas for re-estimation of noise model parameters and clean speech estimation are the same as those in [6], while the updating formulas for convolutional distortion are as follows:

$$\overline{\boldsymbol{\mu}}_{\mathbf{h}} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t) E_{\mathbf{h}}[\mathbf{h}_t|\mathbf{y}_t, m]}{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)} \quad (8)$$

$$\overline{\boldsymbol{\Sigma}}_{\mathbf{h}} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t) E_{\mathbf{h}}[\mathbf{h}_t\mathbf{h}_t^{\top}|\mathbf{y}_t, m]}{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)} - \overline{\boldsymbol{\mu}}_{\mathbf{h}}\overline{\boldsymbol{\mu}}_{\mathbf{h}}^{\top} \quad (9)$$

where $E_{\mathbf{h}}[\mathbf{h}_t|\mathbf{y}_t, m]$ and $E_{\mathbf{h}}[\mathbf{h}_t\mathbf{h}_t^{\top}|\mathbf{y}_t, m]$ are the relevant conditional expectations evaluated as follows:

$$E_{\mathbf{h}}[\mathbf{h}_t|\mathbf{y}_t, m] = \boldsymbol{\mu}_{\mathbf{h}} + \boldsymbol{\Sigma}_{\mathbf{hy},m}\boldsymbol{\Sigma}_{\mathbf{y},m}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},m}) \quad (10)$$

$$E_{\mathbf{h}}[\mathbf{h}_t\mathbf{h}_t^{\top}|\mathbf{y}_t, m] = E_{\mathbf{h}}[\mathbf{h}_t|\mathbf{y}_t, m]E_{\mathbf{h}}^{\top}[\mathbf{h}_t|\mathbf{y}_t, m] + \boldsymbol{\Sigma}_{\mathbf{h}} - \boldsymbol{\Sigma}_{\mathbf{hy},m}\boldsymbol{\Sigma}_{\mathbf{y},m}^{-1}\boldsymbol{\Sigma}_{\mathbf{yh},m} \ . \quad (11)$$

As for initialization, $\boldsymbol{\mu}_{\mathbf{h}}$ is set as a zero vector and $\boldsymbol{\Sigma}_{\mathbf{h}}$ is calculated as

$$\boldsymbol{\Sigma}_{\mathbf{h}} = \alpha\left[\sum_{m=1}^{M}\omega_m\left(\boldsymbol{\Sigma}_{\mathbf{x},m} + \boldsymbol{\mu}_{\mathbf{x},m}\boldsymbol{\mu}_{\mathbf{x},m}^{\top}\right) - \left(\sum_{m=1}^{M}\omega_m\boldsymbol{\mu}_{\mathbf{x},m}\right)\left(\sum_{m=1}^{M}\omega_m\boldsymbol{\mu}_{\mathbf{x},m}\right)^{\top}\right] \quad (12)$$

where $\alpha$ is a control parameter.

To calculate the required statistics for new formulations via high-order VTS approximation, two additional statistics compared with those in [6], namely $\boldsymbol{\Sigma}_{\mathbf{xy}}$ and $\boldsymbol{\Sigma}_{\mathbf{hy}}$ in log-power-spectral domain, should be calculated as

$$\sigma_{xy}^2(i,j) = \sum_{k=0}^{K}\sum_{r=0}^{k}\sum_{p=0}^{k-r} A^j(k,r)C_{k-r}^{p} M_n^j(r)M_h^j(k-r-p)M_x^{ij}(1,p) \quad (13)$$

$$\sigma_{hy}^2(i,j) = \sum_{k=0}^{K}\sum_{r=0}^{k}\sum_{p=0}^{k-r} A^j(k,r)C_{k-r}^{p} M_n^j(r)M_x^j(p)M_h^{ij}(1,k-r-p) \quad (14)$$

where $\sigma_{xy}^2(i,j)$ and $\sigma_{hy}^2(i,j)$ denote the $(i,j)^{\text{th}}$ element of the matrix $\boldsymbol{\Sigma}_{\mathbf{xy}}$ and $\boldsymbol{\Sigma}_{\mathbf{hy}}$, respectively. $C_{k-r}^{p}$ is the number of $p$-combinations from a given set of $k-r$ elements. Other notations are explained in [6].

Furthermore, due to the new assumption of convolutional distortion, the pdf of the variable $\mathbf{z} = \mathbf{x} + \mathbf{h}$ defined in [6], is modified as

$$p(\mathbf{z}_t) = \sum_{m=1}^{M}\omega_m\mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_{\mathbf{x},m} + \boldsymbol{\mu}_{\mathbf{h}}, \boldsymbol{\Sigma}_{\mathbf{x},m} + \boldsymbol{\Sigma}_{\mathbf{h}}) \quad (15)$$

which influences the computation of statistics in [6].

# 3. IVN-based Joint Training of GMM and HMM using VTS-based Feature Compensation

## 3.1. System Overview

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 [9], IVN-based HMM training using VTS-based model compensation is used to train generic HMMs from both clean and noisy speech data. In this work, we propose a novel procedure to perform IVN-based joint training of GMM and HMM 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 trained from multi-condition training data using MFCC features with cepstral mean normalization (CMN).

**Step 2:** *VTS-based feature compensation*

Given the reference GMM, VTS-based feature compensation is applied to each training utterance.

**Step 3:** *Joint training of GMM and HMM*

Based on the compensated features of training set, single pass retraining (SPR) [14] is performed to generate the generic GMM and HMM by using the last updated GMM and HMM 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.

## 3.2. Discussions

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 [9] 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.

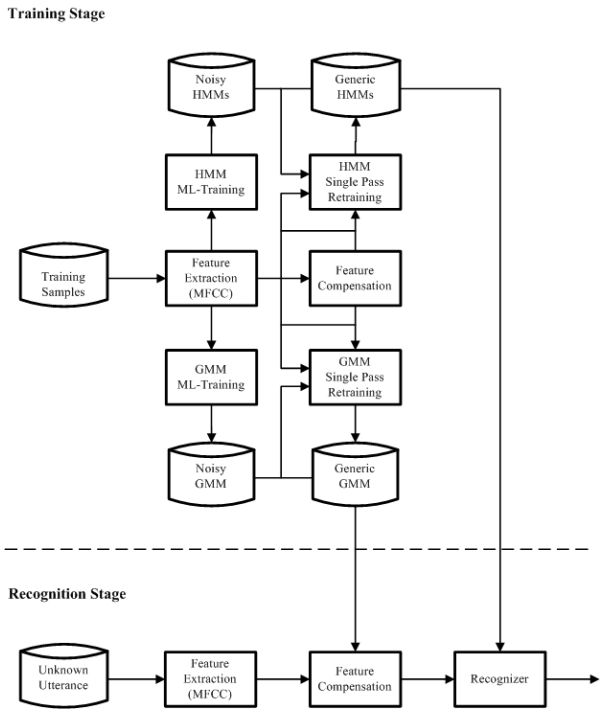


Figure 1: Flowchart of IVN training using VTS-based feature compensation.

# 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 [2, 3, 4, 5].

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. We focus on two "training-testing" conditions for experiments of Aurora3. 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 other one is well-matched (WM) condition, where both training and testing data are recorded by CT and HF microphones. Other control parameters related to our previous work on VTS-based feature compensation can be found in [6]. For boosted mixture learning, the linear scaling factor in sampling boosting is set as suggested in [7] without tuning. For acoustic context expansion, $\Delta$ is set as 3. The control parameter $\alpha$ for initializing the variance of convolutional distortion is set to 1.0. The iteration number $N_{IVN}$ for IVN training is set to 4. 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.

Table 1: Performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using VTS-based feature compensation in the high-mismatch (HM) condition on Aurora3 databases.

| | German | Danish | Finnish | Spanish |
|---|---|---|---|---|
| Baseline | 83.77 | 54.78 | 77.07 | 80.99 |
| VTS-256 | 91.03 | 76.92 | 86.29 | 85.35 |
| VTS-32 | 90.06 | 74.69 | 85.09 | 84.69 |
| +BML | 90.75 | 76.29 | 86.11 | 85.68 |
| +HO | 91.21 | 77.43 | 86.33 | 86.32 |
| +ACE | 91.30 | 77.82 | 89.36 | 87.85 |
| +HSG | 91.31 | 81.70 | 89.52 | 88.54 |

Table 2: Performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using VTS-based feature compensation in the well-matched (WM) condition on Aurora3 databases.

| | German | Danish | Finnish | Spanish |
|---|---|---|---|---|
| Baseline | 92.49 | 90.84 | 93.09 | 93.57 |
| VTS-32 | 93.55 | 91.56 | 93.91 | 93.83 |
| VTS-32-New | 94.01 | 91.77 | 94.12 | 94.36 |
| IVN-Joint | 94.69 | 92.42 | 95.12 | 94.80 |

### 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 in the high-mismatch (HM) condition on Aurora3 databases. VTS-256 refers to the practical solution of feature compensation recommended in [6], namely CMN+VTS(N,H)(MMSE-VTS0) where a first-order VTS approximation in distortion model parameter estimation is used and the mixture number of reference GMM is 256. VTS-32 uses a reference GMM with 32 mixture components to improve the efficiency of feature compensation. +BML uses boosted mixture learning of reference GMM instead of conventional ML training for VTS-32. +HO uses the second order information of VTS approximation for the calculation of noisy speech mean parameters based on +BML. +ACE adds the acoustic context expansion to +HO. +HSG makes the single Gaussian assumption for convolutional distortion based on +ACE. Several observations can be made. First, the performance gap between reference GMMs with 256 and 32 components can be reduced significantly by using BML. Second, using higher order information partially can upgrade the recognition performance. Third, acoustic context expansion is verified to be effective for VTS-based feature compensation. Finally, the single Gaussian assumption for convolutional distortion can be very useful when the convolutional distortion is non-stationary in an utterance, which may be the case for Danish database where the Baseline performance is much lower than those on other databases due to the constant assumption for convolutional distortion in CMN.

Table 2 gives a performance (word accuracy in %) comparison of the baseline system and several robust ASR systems using VTS-based feature compensation in the well-matched (WM) condition on Aurora3 databases. VTS-32 denotes the system where the recommended solution of VTS-based feature compensation in [6] is applied to both training and recognition stages where a 32-component reference GMM trained on original noisy data is used. VTS-32-New is the system where the improved VTS-based feature compensation is used. IVN-Joint represents the system where the proposed SPR-based IVN training with the improved VTS-based feature compensation is used. It is observed that our improved feature compensation approach achieves significant performance improvement compared with the approach in [6], and IVN-Joint system brings additional gains of recognition accuracy for all Aurora3 tasks.

## 5. Conclusion

In this paper, we propose an approach to irrelevant variability normalization based joint training of GMM and HMMs using an improved VTS-based feature compensation for improving the efficiency of feature compensation and upgrading the recognition accuracy of noisy speech. The effectiveness of the proposed approach has been confirmed in an experimental study on Aurora3 tasks.

## 6. References


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

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

[3] 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.

[4] 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.

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

[6] 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.

[7] J. Du and H. Jiang, "Boosted mixture learning of Gaussian mixture hidden Markov models based on maximum likelihood for speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 19, No. 7, pp.2091-2100, 2011.

[8] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *Proc. ICASSP*, 2000, pp.1635-1638.

[9] 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.

[10] 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.

[11] 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.

[12] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," *Proc. ICASSP*, 2005, pp.961-964.

[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.