Unscented Transform with Online Distortion Estimation for HMM Adaptation

Jinyu Li, Dong Yu, Yifan Gong, and Li Deng

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

{jinyli; dongyu; ygong; deng}@microsoft.com

Abstract

In this paper, we propose to improve our previously developed method for joint compensation of additive and convolutive distortions (JAC) applied to model adaptation. The improvement entails replacing the vector Taylor series (VTS) approximation with unscented transform (UT) in formulating both the *static* and *dynamic* model parameter adaptation. Our new JAC-UT method differentiates itself from other UT-based approaches in that it combines the *online* noise and channel distortion estimation and model parameter adaptation in a unified UT framework. Experimental results on the standard Aurora 2 task show that the new algorithm enjoys 20.0% and 16.9% relative word error rate reductions over the previous JAC-VTS algorithm when using the simple and complex backend models, respectively.

Index Terms: unscented transform, vector Taylor series, additive and convolutive distortions, robust ASR, adaptation

1. Introduction

Environment robustness has been one of the most popular research topics in automatic speech recognition (ASR) during past two decades. Techniques tackling robustness issues can be categorized into two classes: feature-domain (e.g., [1][2]) and model-domain (e.g., [3][4]) approaches. Feature-domain approaches enhance the distorted speech features with advanced signal processing methods without adjusting the model parameters while the model-domain approaches adapt the model parameters to make the model better matched to the distorted environment.

In recent years, a model-domain approach that jointly compensates for additive and convolutive distortions (JAC) was proposed and evaluated (e.g., [4][5][6][7][8][9]), yielding promising results. The various JAC-based methods proposed so far use a parsimonious nonlinear *physical* model to describe the environmental distortion and use the vector Taylor series (VTS) approximation technique to find closed-form hidden Markov model (HMM) adaptation and noise/channel parameter estimation formulas. The JAC-VTS model adaptation technique, while achieving noticeable performance improvement over various competing techniques, has the known limitation that the same approximated linear mapping between the clean and distorted speech model parameters is shared across the entire model space even though the true mapping is nonlinear.

In this paper, we propose to address this and related limitations of the JAC-VTS technique by replacing VTS with unscented transformation (UT) in estimating the noise and channel distortions and in adapting the HMM parameters *online*. Originally developed to improve extended Kalman filter, UT [10] is an effective way to estimate mean and variance parameters under nonlinear transformation. It was first introduced to the field of robust ASR in [11]. In that work, the static mean and variance of nonlinearly distorted *speech* signals was estimated using UT, but the authors estimated the static *noise* mean and variance with a simple average of the beginning and ending frames of the current utterance. The technique was improved in [12], where the static noise parameters were estimated online with maximum likelihood estimation (MLE) using the VTS approximation and the estimates were subsequently plugged into the UT formulation to obtain the estimate of the mean and variance of the static distorted speech features. Most recently, Faubel et al. [13] proposed a novel robust feature extraction technique which estimates the parameters of the conditional noise and channel distribution using UT and embeds the estimated parameters into the expectation maximization (EM) [14] framework. Note that in all these approaches [11][12][13], sufficient statistics of only the static features or model parameters are estimated using UT although adaptation of the dynamic model parameters with reliable noise and channel estimations has shown to be important [7].

The JAC-UT approach proposed in this paper differentiates itself from [13] in that it is a model-domain approach while the technique proposed in [13] is a feature-domain one. Our approach also differs from that of [11][12] in that our JAC-UT approach estimates both noise estimation and distorted speech estimation consistently within the same UT framework. Furthermore, our JAC-UT extends the previous work of [11][12][13] by estimating sufficient statistics of not only the static model parameters but also the dynamic model parameters.

We evaluated the JAC-UT technique on the standard Aurora 2 task. The experimental results show that JAC-UT outperforms JAC-VTS by 20.0% and 16.9% in relative word error rate (WER) reductions when using the simple and complex backend models, respectively. The experimental results reported in this paper also shed insight into our earlier work [8][15] on the role of the mixing phase between speech and noise in speech feature enhancement. Specifically, our new results show that with better model space mapping and improved estimation of noise and channel parameters using UT, the performance of a phase-ignored JAC system [8][15] can be significantly improved and the unusually high distortion adjustment term proposed in [8] becomes less important compared with the adjustment introduced under the previous JAC-VTS framework.

The rest of the paper is organized as follows. In Section 2, we describe the novel JAC-UT algorithm. In Section 3, we present the experimental results on the standard Aurora 2 task using both simple and complex back-ends. We summarize our study and conclude the paper in Section 4.

2. JAC-UT Adaptation Algorithms

In this section, we first briefly review the JAC-VTS algorithm and then derive the JAC-UT algorithm for the HMM means and variances on the Mel-frequency cepstral coefficient (MFCC) features for both static and dynamic model parameters. We subsequently describe the algorithm which jointly estimates the additive and convolutive distortion parameters using UT.

2.1. JAC-VTS Adaptation Algorithm

Figure 1 shows a model for degraded speech with both noise (additive) and channel (convolutive) distortions. The observed distorted speech signal $y[m]$ is generated from clean speech $x[m]$ with noise $n[m]$ and channel's impulse response $h[m]$ according to

$$y[m] = x[m] * h[m] + n[m].$$

With discrete Fourier transformation (DFT), the equivalent relationship

$$Y[k] = X[k]H[k] + N[k]$$

can be established in the frequency domain, where k is the frequency-bin index in DFT given a fixed-length time window.

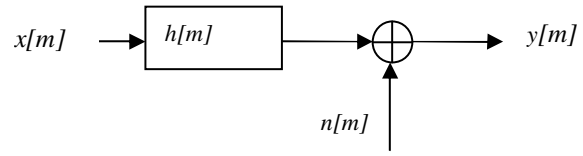

Figure 1: *A model for acoustic environment distortion*

The power spectrum of the distorted speech can then be obtained as

$$|Y[k]|^2 = |X[k]|^2|H[k]|^2 + |N[k]|^2 + 2|X[k]||H[k]||N[k]|\cos\theta_k, \quad (1)$$

where θ_k denotes the (random) angle between the two complex variables $N[k]$ and $(X[k]H[k])$.

It is noted that Eq. (1) is a general formulation for JAC. If $\cos\theta_k$ is set to zero, Eq. (1) becomes

$$|Y[k]|^2 = |X[k]|^2|H[k]|^2 + |N[k]|^2, \quad (2)$$

which is the formulation often used when power spectra [5] are adopted as the acoustic feature. If $\cos\theta_k$ is set to one, we obtain

$$|Y[k]| = |X[k]||H[k]| + |N[k]|, \quad (3)$$

which is the formulation often used when magnitude spectra [7] are adopted as the acoustic feature.

By taking logarithm and multiplying the non-square discrete cosine transform (DCT) matrix C to both sides of Eq. (1) for all the L Mel filter-banks, we obtain the nonlinear distortion model of

$$y = x + h + Clog(1 + \exp(C^{-1}(n - x - h)) + 2\alpha \exp\left(\frac{C^{-1}(n - x - h)}{2}\right), \quad (4)$$

where x, n, h, and y are clean speech, noise, channel, and distorted speech, respectively, in the cepstral domain, and α is a phase related adjustment term. If $\alpha = 0$, Eq. (4) becomes

$$y = x + h + Clog(1 + \exp(C^{-1}(n - x - h)), \quad (5)$$

which is the popular JAC formulation.

Note that $\alpha = 0$ is a reasonable theoretical approximation since this is its mean value and the random value of α is ranged between -1 and 1 in theory [15]. However, it was observed in [8] and [16] that setting $\alpha = 0$ performs much worse than setting $\alpha = 2.5$ using JAC-VTS. A possible explanation is that the noise and channel distortions were estimated with possibly systematic biases since VTS discards the second and higher-order terms. A larger α thus may partially compensate for the biases.

Given its theoretical justification, we assume $\alpha = 0$ and thus use Eq. (5) to describe the feature space distortion hereon. By taking the expectation on both sides of Eq. (5), the static mean value of the distorted speech signal is

$$\mu_y = \mu_x + \mu_h + g(\mu_x, \mu_h, \mu_n) \approx \mu_x + \mu_{h,0} + G(\mu_h - \mu_{h,0}) + (I - G)(\mu_n - \mu_{n,0}), \quad (6)$$

where

$$g(\mu_x, \mu_h, \mu_n) = Clog(1 + \exp(C^{-1}(\mu_n - \mu_x - \mu_h))). \quad (7)$$

By noting,

$$\frac{\partial \mu_y}{\partial \mu_x} = Cdiag\left\{\frac{1}{1 + \exp[C^{-1}(\mu_n - \mu_x - \mu_h]}\right\}C^{-1} = G \quad (8)$$

$$\frac{\partial \mu_y}{\partial \mu_x} = I - G, \quad (9)$$

we can derive the JAC-VTS adaption formulations for the k-th Gaussian in the j-th state as (following [7]):

$$\mu_{y,jk} = \mu_{x,jk} + \mu_h + g(\mu_{x,jk}, \mu_h, \mu_n), \quad (10)$$

$$\Sigma_{y,jk} \approx G(j,k)\Sigma_{x,jk}G(j,k)^T + (I - G(j,k))\Sigma_n(I - G(j,k))^T, \quad (11)$$

$$\mu_{\Delta y,jk} \approx G(j,k)\mu_{\Delta x,jk}, \quad (12)$$

$$\mu_{\Delta\Delta y,jk} \approx G(j,k)\mu_{\Delta\Delta x,jk}, \quad (13)$$

$$\Sigma_{\Delta y,jk} \approx G(j,k)\Sigma_{\Delta x,jk}G(j,k)^T + (I - G(j,k))\Sigma_{\Delta n}(I - G(j,k))^T, \quad (14)$$

$$\Sigma_{\Delta\Delta y,jk} \approx G(j,k)\Sigma_{\Delta\Delta x,jk}G(j,k)^T + (I - G(j,k))\Sigma_{\Delta\Delta n}(I - G(j,k))^T. \quad (15)$$

The online estimation formulas for μ_n, μ_h, Σ_n, $\Sigma_{\Delta n}$, and $\Sigma_{\Delta\Delta n}$ can be found in [9] and are not repeated here.

2.2. Basic UT Algorithm

As in [11], an augmented signal $s = [x^T, n^T]^T$ is formed with a D-dimensional clean speech cepstrum x and a noise cepstrum n, with dimensionality $D_s = D_x + D_n = 2D$.

The UT algorithm samples the augmented signal s with $4D$ sigma points:

$$s_i = \begin{cases} \mu_s + \left(\sqrt{2D\Sigma_s}\right)_i, & \text{if } i = 1 \ldots\ldots 2D \\ \mu_s - \left(\sqrt{2D\Sigma_s}\right)_{i-2D}, & \text{if } i = 2D+1 \ldots\ldots 4D, \end{cases} \quad (16)$$

where μ_s and Σ_s are the mean and covariance of the augmented signal, and $\left(\sqrt{\Sigma}\right)_i$ denotes the i-th column of the square root matrix of Σ.

In the feature space, the transformed sample z_i with a mapping function $f(.)$ is $z_i = f(s_i)$.

In the model space, the mean and variance values are

$$\mu_y = \Sigma_{i=0}^{4D} w_i z_i, \quad (17)$$

$$\Sigma_y = \Sigma_{i=0}^{4D} w_i (z_i - \mu_y)(z_i - \mu_y)^T, \quad (18)$$

where $w_i = 1/4D$ are weights of each sigma point.

2.3. JAC-UT Algorithm

From Eq. (5) the transformed sample z_i for the sigma point s_i is

$$z_i = f(s_i) = f(x_i^T, n_i^T) = x_i + h + Clog(1 + \exp(C^{-1}(n_i - x_i - h))),$$

where $x_i = \mu_x + \delta_{xi}$ and $n_i = \mu_n + \delta_{ni}$, with δ_{xi} and δ_{ni} being the offsets of x_i and n_i from μ_x and μ_n, respectively.

$$\mu_y = \sum_{i=1}^{4D} w_i z_i = \sum w_i \left(\mu_x + \delta_{xi} + \mu_h + C log\left(1 + \exp\left(C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)\right)\right)\right)$$
$$= \sum w_i \mu_x + \sum w_i \delta_{xi} + \sum w_i \mu_h + \sum w_i C\, log\left(1 + \exp\left(C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)\right)\right)$$
$$= \mu_x + \mu_h + \sum w_i C log\left(1 + \exp\left(C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)\right)\right) = \mu_x + \mu_h + g'(\mu_x, \mu_h, \mu_n). \quad (19)$$

$$\frac{\partial \mu_y}{\partial \mu_x} = \frac{\partial \mu_y}{\partial \mu_h}$$
$$= I - \sum w_i C diag\{\exp\left(C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)\right)/(1 + \exp[C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)])\} C^{-1}$$
$$= \sum w_i C diag\{1/(1 + \exp[C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)])\} C^{-1} = G'. \quad (20)$$

$$\mu_n = \mu_{n,0} + \left\{\sum_t \sum_j \sum_k \gamma_t(j,k)\left(I - G'(j,k)\right)^T \Sigma_{y,jk}^{-1}\left(I - G'(j,k)\right)\right\}^{-1}$$
$$\left\{\sum_t \sum_j \sum_k \gamma_t(j,k)\left(I - G'(j,k)\right)^T \Sigma_{y,jk}^{-1}\left(y_t - \mu_{x,jk} - \mu_{h,0} - g'\left(\mu_{n,0}, \mu_{x,jk}, \mu_{h,0}\right)\right)\right\}. \quad (21)$$

$$\mu_h = \mu_{h,0} + \left\{\sum_t \sum_j \sum_k \gamma_t(j,k) G'(j,k)^T \Sigma_{y,jk}^{-1} G'(j,k)\right\}^{-1}$$
$$\left\{\sum_t \sum_j \sum_k \gamma_t(j,k) G'(j,k)^T \Sigma_{y,jk}^{-1}\left(y_t - \mu_{x,jk} - \mu_{h,0} - g'\left(\mu_{n,0}, \mu_{x,jk}, \mu_{h,0}\right)\right)\right\}. \quad (22)$$

They can be easily calculated from Eq. (16).

We thus obtain the static transformed mean values as shown in Eq. (19), where

$$g'(\mu_x, \mu_h, \mu_n) = \Sigma\, w_i C\, log\left(1 + \exp\left(C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)\right)\right). \quad (23)$$

Likewise, the static transformed variance can be calculated with Eq. (18). We can also calculate the derivatives of μ_y with respect to μ_x and μ_h as shown in Eq. (20) and to μ_n as

$$\frac{\partial \mu_y}{\partial \mu_n} = I - G'. \quad (24)$$

EM algorithm is developed in this work as part of the overall JAC-UT algorithm to estimate the noise and channel parameters. Let $\gamma_t(j,k)$ denote the posterior probability for the k-th Gaussian in the j-th state of the HMM, i.e.,

$$\gamma_t(j,k) = p\left(\vartheta_t = j, \epsilon_t = k \middle| Y, \bar{\lambda}\right),$$

where ϑ_t denotes the state index, and ϵ_t denotes the Gaussian index at time frame t. $\bar{\lambda}$ is the old parameter set of noise and channel. Embedding μ_y into the EM auxiliary function, and taking the first derivative with respect to μ_n and μ_h,we obtain

$$\frac{\partial Q}{\partial \mu_n} \sim \sum_t \sum_j \sum_k \gamma_t(j,k)\left(I - G'(j,k)\right)^T \Sigma_{y,jk}^{-1}\left(y_t - \mu_{y,jk}\right) = 0, \quad \text{or}$$
$$\frac{\partial Q}{\partial \mu_h} \sim \sum_t \sum_j \sum_k \gamma_t(j,k) G'(j,k)^T \Sigma_{y,jk}^{-1}\left(y_t - \mu_{y,jk}\right) = 0.$$

Because μ_y is a nonlinear function of μ_n and μ_h, by linearlizing it as

$$\mu_y = \mu_x + \mu_{h,0} + G'^{(\mu_h - \mu_{h,0})} + (I - G')\left(\mu_n - \mu_{n,0}\right) \quad (25)$$

we obtain the closed-form solution as shown in Eqs. (21) and (22).

Comparing Eqs. (21) and (22) with the solution in [9] where VTS is used, we can see that the solution formulas are the same except we are using weighted sums $G'(j,k)$ (defined in Eq. (20)) and $g'\left(\mu_{x,jk}, \mu_{h,0}, \mu_{n,0}\right)$ (defined in Eq. (23)) to replace $G(j,k)$ (defined in Eq. (8)) and $g\left(\mu_{x,jk}, \mu_{h,0}, \mu_{n,0}\right)$ (defined in Eq. (7)).

To estimate the dynamic parameters for distorted speech, linearization is still needed as discussed in [9]. Inferring from Eq. (25) and Eq. (6), we can similarly use $G'(j,k)$ to replace $G(j,k)$ in Eqs. (12), (13), (14), and (15), and obtain the corresponding dynamic model formulations for the distorted speech signal. The re-estimation formulas for the dynamic noise variances are the same as that in [9] because the adaptation formulations share the same formulas.

3. Experimental Evaluation

The proposed JAC-UT algorithm presented in Section 2 has been evaluated on the standard Aurora 2 task [17] of recognizing digit strings in noise and channel distorted environments. The clean training set is used to train the baseline maximum likelihood estimation (MLE) HMMs. The test material consists of three sets of distorted utterances. Set-A and set-B contain eight different types of additive noise while set-C contains two different types of noise and additional channel distortion. The baseline experiment setup follows the standard script provided by ETSI, including the standard simple and complex backend [1] of HMMs trained using the HTK toolkit.

The features are 13-dimension MFCCs, appended by their first- and second-order time derivatives. The cepstral coefficient of order zero is used instead of the log energy in the original script. We use power spectra for MFCC extraction in all experiments.

The JAC-UT algorithm presented in this paper is used to adapt the ML-trained HMMs utterance by utterance for the entire test set (Sets-A, B, and C). The implementation steps described in [7] are used in the experiments. We use the first and last 20 frames from each utterance for initializing the noise means and variances. Only one-pass processing is used in the reported experiments.

Table 1: *Recognition accuracies (Acc) under the baseline, JAC-VTS, and different JAC-UT setups for clean-trained simple backend HMMs. Power spectra are used to extract MFCC features.*

Setup	Acc
Baseline	58.70%
JAC-VTS	88.35%
Static noise/channel estimated in VTS, static model mean/variance updated in UT	89.21%
Static noise/channel estimated in UT, static model mean/variance updated in UT	89.34%
All estimates/updates are in UT	90.68%

To examine the contribution of individual components in the JAC-UT algorithm, we conducted experiments using the

$$G'' = I - \Sigma w_i C diag\left\{\frac{\exp\left(C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)\right) + \alpha \exp[C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)/2\ ]}{1 + \exp[C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)] + 2\alpha \exp[C^{-1}(\mu_n + \delta_{ni} - \mu_x - \delta_{xi} - \mu_h)/2\ ]}\right\} C^{-1} \quad (26)$$

JAC-VTS setting, and then gradually switched components from VTS to UT formulation. As shown in Table 1, the baseline accuracy (Acc) is 58.70% using the clean-trained simple backend model. When adapting with the normal JAC-VTS (i.e., $\alpha = 0$ in phase-JAC-VTS, all noise/channel parameters are online estimated), the Acc improves to 88.35%. If we use VTS to estimate the static noise and channel means and then plug them into Eqs. (17) and (18) to adapt the static model mean and variance as done in [12], the Acc is increased to 89.21%. After applying Eqs. (21) and (22) to estimate the noise and channel means, the Acc further improves to 89.34%. Finally, the dynamic model parameters are updated by replacing the VTS-derived $G(j,k)$ with the UT-derived $G'(j,k)$ in Eqs. (12)-(15), and the dynamic noise variances are estimated online. This setting achieves the highest accuracy of 90.68%, which translates to a 20.0% relative WER reduction over the normal JAC-VTS algorithm. This demonstrates that the normal JAC method (without any phase term) can have better performance with an improved estimate of model space mapping using UT.

In Table 2, we show experimental results using the complex backend with the JAC-UT model adaptation technique. When $\alpha = 0$, JAC-UT obtains 91.68% Acc, which stands for 16.9% relative WER reduction from the 89.99% Acc achieved using the JAC-VTS approach. Note that this accuracy is still lower than the 93.32% Acc achieved in [8] with phase-adjusted JAC-VTS when $\alpha = 2.5$.

Table 2: *Recognition accuracies (Acc) under the settings of baseline, phase-JAC-VTS, and alpha-JAC-UT with different α for clean-trained complex backend HMMs. Power spectra are used to extract MFCC features.*

Settings	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2.5$
phase-JAC-VTS	89.99%	91.85%	92.70%	93.32%
alpha-JAC-UT	91.68%	92.57%	92.91%	93.30%

In the formulation of JAC-UT, linearization is still used in order to achieve a closed-form solution. As argued in [9], a large value of α may be used to compensate for the linearization bias. Therefore, we try to keep the UT model space mapping in Eqs. (17) and (18), and use the G'' defined in Eq. (26) to replace G' defined in Eq. (20) by introducing an α term with each element similar to the format in [9]. Note that $G'' = G'$ when $\alpha = 0$. We call this method alpha-JAC-UT instead of phase-JAC-UT because there is no phase term in this feature space distortion model and the α term is only used to compensate for the linearization bias.

The results in Table 2 demonstrate that with larger α values, JAC-UT can further improve the accuracy. When α equals 0, 0.5, and 1, alpha-JAC-UT outperforms phase-JAC-VTS with reduced relative gains as α is increased. When $\alpha = 2.5$, these two methods obtain almost the same accuracy.

4. Conclusions

In this paper, we have presented our recent development of the JAC-UT algorithm for HMM adaptation and demonstrated its effectiveness on the standard Aurora 2 environment-robust ASR task. This approach unifies the static and dynamic model parameter adaptation with online estimation of noise and channel parameters in the UT framework, distinguishing itself from prior arts.

In the experimental evaluation using the standard Aurora 2 task, the proposed JAC-UT algorithm has achieved 20.0% and 16.9% relative WER reduction from JAC-VTS algorithm, with the clean-trained simple and complex HMM backends, respectively. The UT formulation and the experimental results shed light onto the previous unsatisfactory performance with $\alpha = 0$ using the phase-JAC-VTS technique. We conclude from this work that JAC methods can obtain more satisfactory accuracy by utilizing a better model space mapping.

To obtain a closed-form solution in this work, we still retain the linearization step in the JAC-UT framework. Alpha-JAC-UT is used to boost the accuracy by adding an α term to compensate for the linearization loss. This partially exposes the weakness of our current JAC-UT formulation. Our future work involves further improvement of the performance of JAC-UT without employing linearization. Note that UT brings more computation costs than VTS. It is important to reduce the costs in the future.

5. References

[1] Macho, D., et al., "Evaluation of a noise-robust DSR front-end on Aurora databases," *Proc. ICSLP*, pp. 17–20, 2002.

[2] Yu, D., et al., "Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor", IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 5, pp. 1061-1070, July 2008

[3] Gales, M. J. F. and Young, S., "An improved approach to the hidden Markov model decomposition of speech and noise," *Proc. ICASSP*, Vol. I, pp. 233–236, 1992.

[4] Gong, Y., "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Trans. Speech and Audio Proc.*, Vol. 13, No. 5, pp. 975-983, 2005.

[5] Moreno, P., *Speech Recognition in Noisy Environments*. PhD. Thesis, Carnegie Mellon University, 1996.

[6] Liao, H. and Gales, M. J. F., "Joint uncertainty decoding for robust large vocabulary speech recognition," *Tech. Rep. CUED/TR552*, University of Cambridge, 2006.

[7] Li, J., et al., "High-performance HMM adaptation with joint compensation of additive and convolutive distortions," *Proc. IEEE ASRU*, 2007.

[8] Li, J., et al., "HMM adaptation using a phase-sensitive acoustic distortion model for environment-robust speech recognition," *Proc. IEEE ICASSP*, 2008.

[9] Li, J., et al., "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," Computer Speech and Language, no. 3, vol. 23, Elsevier, 2009.

[10] Julier, S.J. and Uhlmann, J.K., "Unscented filtering and nonlinear estimation," *Proceedings of IEEE*, vol. 92, no. 3, pp. 401-422, 2004

[11] Hu, Y. and Huo, Q., "An HMM compensation approach using unscented transformation for noisy speech recognition," *Proc. ISCSLP*, 2006.

[12] Xu, H. and Chin, K.K., "Comparison of estimation techniques in joint uncertainty decoding for noise robust speech recognition," *Proc. Interspeech*, 2009.

[13] Faubel, F., McDonough, J., and Klakow, D., "On expectation maximization based channel and noise estimation beyond the vector Taylor series expansion," in *Proc. ICASSP*, pp. 4294-4297, 2010.

[14] Dempster, A., Laird, N., and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B, 39(1), pp. 1–38, 1977.

[15] Deng, L., Droppo, J., and Acero, A., "Enhancement of log-spectra of speech using a phase-sensitive model of the acoustic environment," *IEEE Trans. Speech and Audio Proc.*, Vol. 12, No. 3, pp. 133-143, 2004.

[16] Gales, M. J. F. and Flego, F., "Discriminative classifiers and generative kernels for noise robust speech recognition," *Technical Report*, CUED, Cambridge, 2008.

[17] Hirsch, H.G. and Pearce, D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. ISCA ITRW ASR*, 2000.