Joint Uncertainty Decoding with Unscented Transform for Noise Robust Subspace Gaussian Mixture Models

Liang Lu, Arnab Ghoshal, and Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
{liang.lu, a.ghoshal, s.renals}@ed.ac.uk

Abstract

Common noise compensation techniques use vector Taylor series (VTS) to approximate the mismatch function. Recent work shows that the approximation accuracy may be improved by sampling. One such sampling technique is the unscented transform (UT), which draws samples deterministically from clean speech and noise model to derive the noise corrupted speech parameters. This paper applies UT to noise compensation of the subspace Gaussian mixture model (SGMM). Since UT requires relatively smaller number of samples for accurate estimation, it has significantly lower computational cost compared to other random sampling techniques. However, the number of surface Gaussians in an SGMM is typically very large, making the direct application of UT, for compensating individual Gaussian components, computationally impractical. In this paper, we avoid the computational burden by employing UT in the framework of joint uncertainty decoding (JUD), which groups all the Gaussian components into small number of classes, sharing the compensation parameters by class. We evaluate the JUD-UT technique for an SGMM system using the Aurora 4 corpus. Experimental results indicate that UT can lead to increased accuracy compared to VTS approximation if the JUD phase factor is untuned, and to similar accuracy if the phase factor is tuned empirically.

1. Introduction

The accuracy of speech recognisers normally degrades significantly in noisy environments, which limits their deployment in many real world applications. One of the main barriers to improving noise robustness lies in the highly nonlinear mismatch function between clean and noise corrupted speech, which is challenging for either de-noising in feature domain or compensation in model domain. Approaches to overcome this include approximation techniques such as vector Taylor series (VTS) [1, 2] which approximate the mismatch function by truncated vector Taylor series expansions or sampling techniques such as data-driven parallel model combination (DPMC) [3] which draw samples from clean speech and noise models to synthesise noisy speech. VTS is relatively computationally efficient, as a closed form solution can be obtained. However, it may lead to biased estimates since it neglects the higher order coefficients. DPMC, on the other hand, is able to estimate the distribution of noisy speech with sufficient accuracy by drawing a large number of samples, but with heavy computational cost, making it infeasible for large vocabulary speech recognition.

Recently, the unscented transform (UT) [4] has been applied to noise compensation in both feature and model domains [5–8], and has achieved good results. Unlike DPMC, UT draws samples deterministically from the *sigma points*—a set of points chosen to have the same mean and covariance as the original distribution. In UT it assumed that the mean and covariance of the nonlinear system can be derived from sigma points [4], although a recent review [9] pointed out that this is not guaranteed depending on the nonlinear system and parameterisation of UT. Based on GMM system settings, UT can result in a more accurate estimate compared to first-order VTS, while its computational cost is much lower than DPMC [7, 8].

In this paper, we apply UT to compensate an SGMM acoustic model [10] against noise. Compared to conventional acoustic modelling based on GMMs, SGMMs construct a much larger number of Gaussian components, which makes compensating each component directly computationally infeasible. To address this, we apply UT in the joint uncertainty decoding (JUD) framework [11], clustering the set of Gaussian components into a smaller number of classes [7], with the compensation parameters being shared by the components belonging to the same class. This greatly reduces the computational cost without notably sacrificing the accuracy. We evaluate this approach on the Aurora 4 task, and observe that UT can successfully compensate the SGMM acoustic model within the JUD framework and lead to higher accuracy compared to VTS compensation.

The rest of the paper is organised as follows. We first review JUD in section 2, and then discuss VTS and UT noise compensation in the JUD framework. Section 3 applies JUD to SGMM acoustic model and we report experimental results in section 4 followed by conclusion in section 5.

2. Joint Uncertainty Decoding

If we denote the clean and noisy speech observations as $\mathbf{x}$ and $\mathbf{y}$, respectively, then, for Gaussian component m, joint uncertainty decoding (JUD) [11] approximates the likelihood of noisy observations given the model as:

$$p(\mathbf{y} \mid m) = \int p(\mathbf{y} \mid \mathbf{x}, m) p(\mathbf{x} \mid m) d\mathbf{x} \approx \int p(\mathbf{y} \mid \mathbf{x}, r_m) p(\mathbf{x} \mid r_m) d\mathbf{x}, \quad (1)$$

where $\mathbf{x}$ is viewed as a latent variable and r_m denotes the regression class that m belongs to. The conditional probability $p(\mathbf{y} \mid \mathbf{x}, m)$ indicates the effect of noise on clean speech for Gaussian component m. In JUD, the exact conditional probability $p(\mathbf{y} \mid \mathbf{x}, m)$ is approximated by $p(\mathbf{y} \mid \mathbf{x}, r_m)$, significantly reducing the computational cost if the number of r is much smaller than that of m.

The conditional distribution $p(\mathbf{y} \mid \mathbf{x}, r_m)$ is derived from the joint distribution of clean and noise corrupted speech which is assumed to be Gaussian. For the rth regression class

$$p\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \mid r\right) \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x^{(r)} \\ \boldsymbol{\mu}_y^{(r)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_x^{(r)} & \boldsymbol{\Sigma}_{yx}^{(r)} \\ \boldsymbol{\Sigma}_{xy}^{(r)} & \boldsymbol{\Sigma}_y^{(r)} \end{bmatrix}\right), \quad (2)$$

By marginalising the likelihood in (1), the likelihood of corrupted speech for the m_{th} component can be approximated as:

$$p(\mathbf{y} \mid m) \approx |\mathbf{A}^{(r)}| \mathcal{N}\left(\mathbf{A}^{(r)}\mathbf{y} + \mathbf{b}^{(r)}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_b^{(r)}\right). \quad (3)$$

where the JUD transformation parameters are obtained as:

$$\mathbf{A}^{(r)} = \boldsymbol{\Sigma}_x^{(r)} \boldsymbol{\Sigma}_{yx}^{(r)-1} \quad (4)$$

$$\mathbf{b}^{(r)} = \boldsymbol{\mu}_x^{(r)} - \mathbf{A}^{(r)} \boldsymbol{\mu}_y^{(r)} \quad (5)$$

$$\boldsymbol{\Sigma}_b^{(r)} = \mathbf{A}^r \boldsymbol{\Sigma}_y^{(r)} \mathbf{A}^{(r)T} - \boldsymbol{\Sigma}_x^{(r)} \quad (6)$$

The transformation parameters, $\boldsymbol{\mu}_x^{(r)}$ and $\boldsymbol{\Sigma}_x^{(r)}$, can be estimated from the clean speech model using a regression tree. $\boldsymbol{\mu}_y^{(r)}$, $\boldsymbol{\Sigma}_y^{(r)}$ and the cross covariance $\boldsymbol{\Sigma}_{yx}^{(r)}$ can be obtained by the following mismatch function:

$$\begin{aligned} \mathbf{y_s} &= \mathbf{x}_s + \mathbf{h}_s + \mathbf{C} \log\Big[\mathbf{1} + \exp\left(\mathbf{C}^{-1}(\mathbf{n}_s - \mathbf{x}_s - \mathbf{h}_s)\right) \\ &\quad + 2\boldsymbol{\alpha} \bullet \exp\left(\mathbf{C}^{-1}(\mathbf{n}_s - \mathbf{x}_s - \mathbf{h}_s)/2\right)\Big] \\ &= f(\mathbf{x}_s, \mathbf{n}_s, \mathbf{h}_s, \boldsymbol{\alpha}), \end{aligned} \quad (7)$$

where the subscript s denotes the static parameters, and $\mathbf{1}$ is the unit vector. Here, $\log(\cdot)$, $\exp(\cdot)$ and $\bullet$ denote the element-wise logarithm, exponentiation and multiplication. $\mathbf{n}_s$ and $\mathbf{h}_s$ are static additive and convolutional noise, respectively. $\mathbf{C}$ is the truncated discrete cosine transform (DCT) matrix, and $\mathbf{C}^{-1}$ indicates its pseudoinverse. $\boldsymbol{\alpha}$ denotes the phase factor [12, 13].

2.1. Joint uncertainty decoding with VTS

As in standard noise compensation, in equation (7) additive noise is modelled by a single Gaussian $\mathbf{n} \sim \mathcal{N}(\boldsymbol{u}_n, \boldsymbol{\Sigma}_n)$, and the convolutional noise is assumed to be constant $\mathbf{h} = \boldsymbol{\mu}_h$. The mismatch function is highly nonlinear, which makes it difficult to derive the parameters for the noise corrupted speech $\mathbf{y}$. We may use a first order VTS approximation [1] to linearise the mismatch function around the expansion point $\{\boldsymbol{\mu}_{xs}^{(r)}, \boldsymbol{\mu}_{hs}, \boldsymbol{\mu}_{ns}\}$, which results in:

$$\begin{aligned} \mathbf{y}_s \mid r \approx f(\boldsymbol{\mu}_{xs}^r, \boldsymbol{\mu}_{hs}, \boldsymbol{\mu}_{ns}, \boldsymbol{\alpha}) + \mathbf{G}_x^{(r)}\left(\mathbf{x}_s - \boldsymbol{\mu}_{xs}^{(r)}\right) \\ + \mathbf{G}_n^{(r)}\left(\mathbf{n}_s - \boldsymbol{\mu}_{ns}\right). \end{aligned} \quad (8)$$

$\mathbf{G}_x^{(r)}$ and $\mathbf{G}_n^{(r)}$ denote the Jacobian matrices:

$$\mathbf{G}_x^{(r)} = \frac{\partial f(\cdot)}{\partial \mathbf{x}_s}\Big|_{\boldsymbol{\mu}_{xs}^{(r)}, \boldsymbol{\mu}_{hs}, \boldsymbol{\mu}_{ns}}, \quad \mathbf{G}_n^{(r)} = \mathbf{I} - \mathbf{G}_x^{(r)}. \quad (9)$$

By taking expectations, we can obtain

$$\boldsymbol{\mu}_{ys}^{(r)} = f(\boldsymbol{\mu}_{xs}^r, \boldsymbol{\mu}_{hs}, \boldsymbol{\mu}_{ns}, \boldsymbol{\alpha}) \quad (10)$$

$$\boldsymbol{\Sigma}_{ys}^{(r)} = \mathbf{G}_x^{(r)} \boldsymbol{\Sigma}_{xs} \mathbf{G}_x^{(r)T} + \mathbf{G}_n^{(r)} \boldsymbol{\Sigma}_{ns} \mathbf{G}_n^{(r)T} \quad (11)$$

$$\boldsymbol{\Sigma}_{ysxs}^{(r)} = \mathbf{G}_x^{(r)} \boldsymbol{\Sigma}_{xs} \quad (12)$$

The dynamic parameters can be derived from a continuous time approximation [14]:

$$\begin{aligned} \boldsymbol{\Delta}\mathbf{y}_t \approx \frac{\partial \mathbf{y}}{\partial t}|_t &= \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial t}|_t + \frac{\partial \mathbf{y}}{\partial \mathbf{n}} \frac{\partial \mathbf{n}}{\partial t}|_t \\ &\approx \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \boldsymbol{\Delta}\mathbf{x}_t + \frac{\partial \mathbf{y}}{\partial \mathbf{n}} \boldsymbol{\Delta}\mathbf{n}_t \end{aligned} \quad (13)$$

Hence, by taking expectations we obtain:

$$\boldsymbol{\mu}_{\Delta y}^{(r)} \approx \mathbf{G}_x^{(r)} \boldsymbol{\mu}_{\Delta x}^{(r)} \quad (14)$$

$$\boldsymbol{\Sigma}_{\Delta y}^{(r)} \approx \mathbf{G}_x^{(r)} \boldsymbol{\Sigma}_{\Delta x}^{(r)} \mathbf{G}_x^{(r)T} + \mathbf{G}_n^{(r)} \boldsymbol{\Sigma}_{\Delta n} \mathbf{G}_n^{(r)T}. \quad (15)$$

Similar expressions can be obtained for delta-delta coefficients. For the cross covariance, we can also obtain

$$\boldsymbol{\Sigma}_{\Delta y \Delta x}^{(r)} \approx \mathbf{G}_x^{(r)} \boldsymbol{\Sigma}_{\Delta x}^{(r)}. \quad (16)$$

Here we have assumed $\mathbb{E}[\boldsymbol{\Delta}\mathbf{n}] = \mathbf{0}$. The non-zero assumption of the dynamic coefficients for additive noise mean has been investigated [13], but the results indicate that it does not lead to improvement when compensating the clean speech variance.

2.2. Joint uncertainty decoding with UT

Unlike VTS which approximates the nonlinear function by a linear function to estimate the distribution of $\mathbf{y}$, sampling approaches draw samples from the distributions of $\mathbf{x}$ and $\mathbf{n}$ to synthesise noisy samples from which to estimate its distribution[1]. UT is a deterministic sampling

[1] We don't draw samples for $\mathbf{h}$ because we assume its distribution is a delta function as in section 2.1.

approach. Let $\mathbf{z} = \begin{bmatrix} \mathbf{x}_s \\ \mathbf{n}_s \end{bmatrix}$ be the combined vector, then UT draws samples as

$$\mathbf{z}_0^{(r)} = \boldsymbol{\mu}_z^{(r)}, \tag{17}$$

$$\mathbf{z}_i^{(r)} = \boldsymbol{\mu}_z^{(r)} + \left[\sqrt{(2d+\kappa)\boldsymbol{\Sigma}_z^{(r)}}\right]_i, \tag{18}$$

$$\mathbf{z}_{i+d}^{(r)} = \boldsymbol{\mu}_z^{(r)} - \left[\sqrt{(2d+\kappa)\boldsymbol{\Sigma}_z^{(r)}}\right]_i, \tag{19}$$

where $i = 1, \ldots, d$, and $\sqrt{\mathbf{A}}$ and $[\mathbf{A}]_i$ denote the Cholesky decomposition and i_{th} column of the matrix $\mathbf{A}$ respectively. κ is a tuning parameter, d is the dimensionality of $\mathbf{z}$, and

$$\boldsymbol{\mu}_z^{(r)} = \begin{bmatrix} \boldsymbol{\mu}_{xs}^{(r)} \\ \boldsymbol{\mu}_{ns} \end{bmatrix}, \quad \boldsymbol{\Sigma}_z^{(r)} = \begin{bmatrix} \boldsymbol{\Sigma}_{xs}^{(r)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ns} \end{bmatrix}. \tag{20}$$

After obtaining the noise and clean speech samples $\{\mathbf{n}_0, \ldots, \mathbf{n}_{2d}\}$ and $\{\mathbf{x}_0, \ldots, \mathbf{x}_{2d}\}$, the noise corrupted speech samples $\{\mathbf{y}_0, \ldots, \mathbf{y}_{2d}\}$ can be derived by the mismatch function (7) and the static parameters can be obtained by

$$\boldsymbol{\mu}_{ys}^{(r)} = \sum_{i=0}^{2d} w_i \mathbf{y}_i \tag{21}$$

$$\boldsymbol{\Sigma}_{ys}^{(r)} = \sum_{i=0}^{2d} w_i \mathbf{y}_i \mathbf{y}_i^T - \boldsymbol{\mu}_{ys}\boldsymbol{\mu}_{ys}^T, \tag{22}$$

$$\boldsymbol{\Sigma}_{ysxs}^{(r)} = \sum_{i=0}^{2d} w_i \mathbf{y}_i \mathbf{x}_i^T - \boldsymbol{\mu}_{ys}\boldsymbol{\mu}_{xs}^T, \tag{23}$$

where the weights are defined in UT as

$$w_0 = \frac{\kappa}{d+\kappa}, \quad w_i = \frac{1}{2(d+\kappa)}. \tag{24}$$

In this work, we set $\kappa = 1/2$ to give the equal weight to all the samples [4] . For the dynamic coefficients, we still use the continuous time approximation which requires linearisation as VTS. Unlike equation (9), the Jacobian is obtained by all the samples rather than just the mean as

$$\tilde{\mathbf{G}}_x^{(r)} = \sum_{i=0}^{2d} w_i \frac{\partial f(\cdot)}{\partial \mathbf{x}_{is}}|_{\mathbf{z}_{is}, \boldsymbol{\mu}_{hs}}, \quad \tilde{\mathbf{G}}_n^{(r)} = \mathbf{I} - \tilde{\mathbf{G}}_x^{(r)} \tag{25}$$

In this work, however, we found that using the Jacobian (25) to linearise the static covariance $\boldsymbol{\Sigma}_{ys}$ and $\boldsymbol{\Sigma}_{ysxs}$ can achieve better results, as the static and dynamic coefficients are derived in a consistent fashion.

3. Joint Uncertainty Decoding for SGMMs

In the SGMM acoustic model [10], the HMM state is modelled as:

$$P(\mathbf{y}_t \mid j) = \sum_{k=1}^{K_j} c_{jk} \sum_{i=1}^{I} w_{jki} \mathcal{N}(\mathbf{y}_t \mid \boldsymbol{\mu}_{jki}, \boldsymbol{\Sigma}_i) \tag{26}$$

$$\boldsymbol{\mu}_{jki} = \mathbf{M}_i \mathbf{v}_{jk} \tag{27}$$

$$w_{jki} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jk}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jk}} \tag{28}$$

where t denotes the time frame, j the HMM state index, k the sub-state index [10], I the number of Gaussians, and K_j the number of sub-states in state j. c_{jk} is a sub-state mixture coefficient and $\boldsymbol{\Sigma}_i$ is the i-th covariance matrix. $\mathbf{v}_{jk} \in \mathbb{R}^S$ is referred to as the sub-state vector, where S denotes the subspace dimension. The matrices $\mathbf{M}_i$ and the vectors $\mathbf{w}_i$ span the model subspaces for Gaussian means and weights respectively, and are used to derive the GMM parameters given sub-state vectors (equations (27) and (28)). As the number of Gaussians is very large. a universal background model (UBM) is also introduced, which is a mixture of full covariance Gaussians of size I, to initialise the model and prune the Gaussians for both training and decoding.

In [15], we have shown that using UBM as the regression model for JUD works well for SGMM acoustic models in terms of both accuracy and computational cost. By this approach, given the JUD transforms the likelihood becomes:

$$P(\mathbf{y}_t \mid j, \mathcal{M}_n) = \sum_{k=1}^{K_j} c_{jk} \sum_{i=1}^{I} w_{jki} \left|\mathbf{A}^{(i)}\right| \times \mathcal{N}\left(\mathbf{A}^{(i)}\mathbf{y}_t + \mathbf{b}^{(i)}; \boldsymbol{\mu}_{jki}, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_b^{(i)}\right) \tag{29}$$

where $\mathbf{A}^{(i)}, \mathbf{b}^{(i)}$ and $\boldsymbol{\Sigma}_b^{(i)}$ are derived from the i_{th} Gaussian in the UBM together with the noise model. $\mathcal{M}_n$ denotes the noise model as $\mathcal{M}_n = \{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\mu}_h\}$. To estimate the noise model $\mathcal{M}_n$, we applied the gradient based approach for JUD-UT system to optimise the auxiliary function as

$$\mathcal{Q}(\mathcal{M}_n) = \sum_{jkit} \gamma_{jki}(t) \Bigg[\log \left|\mathbf{A}^{(i)}\right| + \log \mathcal{N}\left(\mathbf{A}^{(i)}\mathbf{y}_t + \mathbf{b}^{(i)}; \boldsymbol{\mu}_{jki}, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_b^{(i)}\right) \Bigg] \tag{30}$$

where $\gamma_{jki}(t) = p(j, k, i \mid \mathbf{y}_t)$ is the Gaussian component posterior. The same was also used for the JUD-VTS system in [15].

Table 1: WERs of noise compensation by JUD on SGMM systems with $\boldsymbol{\alpha} = \mathbf{0}$.

Methods	A	B	C	D	Avg
Clean model	5.2	58.2	50.7	72.1	**59.9**
MTR model	6.8	15.2	18.6	32.3	**22.2**
JUD-VTS init	5.3	22.5	36.8	47.4	32.9
+1st iter	5.1	15.8	24.6	33.8	23.4
+2nd iter	5.1	15.0	19.8	29.7	**20.9**
+UT re-est	5.0	14.0	20.7	28.4	**20.0**
JUD-UT init	5.2	19.8	36.9	44.7	30.6
+1st iter	4.9	15.0	23.4	30.6	21.6
+2nd iter	4.9	14.3	18.4	26.9	**19.3**

4. Experiments

We evaluated the JUD-UT noise compensation for SGMMs on the Aurora 4 corpus, which is derived from the Wall Street Journal (WSJ0) 5k-word closed vocabulary transcription task. The clean training set contains about 15 hours audio, and Aurora 4 provides a noisy version, which allows multi-condition training (MTR). The test set has 300 utterances from 8 speakers. The first test set "`test01`" (set A) was recorded using a close talking microphone, similar to the clean training data. "`test02`" to "`test07`" (set B) were obtained by adding six different types of noise, with randomly selected SNRs ranging from 5dB to 15dB to set A. "`test08`" (set C) was recording using a desk-mounted secondary microphone and the same type of noise was added to this set which gives "`test09`" to "`test14`" (set D). In the following experiments, we used 39 dimensional feature vectors comprising 12th order mel frequency cepstral coefficients (MFCCs), and their first and second derivatives. We used the standard WSJ 5k bigram language model.

The systems were built using the Kaldi speech recognition toolkit [16]. We used $I = 400$ components in the UBM and a subspace dimension $S = 40$ in the SGMM-based systems. There were about 3,900 tied triphone states, and about 16,000 substates were used in total, resulting in 6.4 million surface Gaussians. Unlike [15], we did not split the speech and silence in the UBM model as we found the gains disappear after introducing the phase term. Table 1 gives the baseline results using clean and MTR models. Compared to published results on this task using GMM acoustic model, the SGMM system has a lower WER than the GMM system matched condition, but not for mismatched condition.

4.1. Comparison of JUD-UT and JUD-VTS system

In the framework of JUD, we compared the performance of UT and VTS for noise compensation of the SGMM acoustic model. In these experiments, we initialised the noise model by the first and last 20 frames of each utterance, and the results are shown by "JUD-VTS init" and "JUD-UT init". We then updated the noise model by either UT or VTS using the hypothesis from the previous decoding results. Here, we did not use the phase term, i.e. $\boldsymbol{\alpha} = 0$. The results are shown in Table 1 in which, after two decoding passes, the JUD-VTS system achieves the average WER of 20.9%, indicated by JUD-VTS "+2nd iter". Given these noise model, we re-estimate the JUD compensation parameters using UT and can reduce the WER to be 20.0%. This shows that UT can lead to more accurate estimate in this condition given the same noise model compared to VTS. If we update the noise model from scratch, we achieve 19.3% WER after two decoding passes, which is considerably better than that of 20.9% for JUD-VTS system, and also 22.2% of MTR baseline.

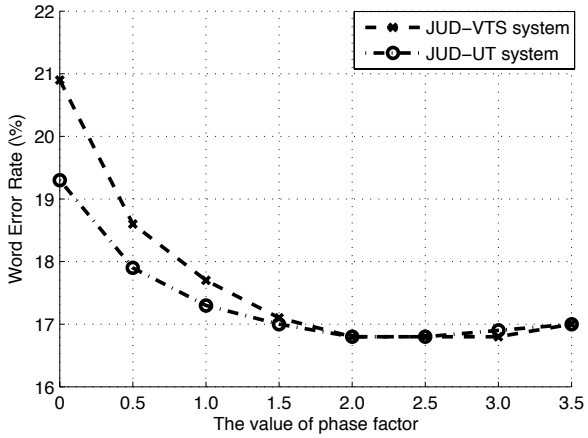

Figure 1: Average WER with respect to the phase term $\boldsymbol{\alpha}$ for JUD with VTS and UT compensation for SGMM systems. They achieve almost the same accuracy after increasing the value of phase term.

We have previously shown that the non-zero phase term can significantly affect the JUD with VTS compensation for SGMM acoustic models [15]. Here, we investigate the effect of the phase term for the UT system. Again, we do not estimate the value of $\boldsymbol{\alpha}$ (as in [12]) but set all the coefficients of $\boldsymbol{\alpha}$ empirically to be a fixed value [13]. Figure 1 graphs the average WERs. Similar to the JUD-VTS system and consistent with the observations in [8], the phase factor also affects JUD with UT system, and after increasing the value of $\boldsymbol{\alpha}$, the gap between JUD-VTS and JUD-UT system shrinks, and both system achieve the same lowest WER, 16.8%, when $\boldsymbol{\alpha} = \mathbf{2.0}$. A non-zero value of $\boldsymbol{\alpha}$ is able to compensate for the linearisation bias [8].

4.2. Analysis of the effect of phase factors

To gain further insight to the effect of phase term, we calculated the total variance of $\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_b^{(i)}$ and averaged it by I and the number of test utterances. The plot is shown in Figure 2 for JUD-UT system. The average value of

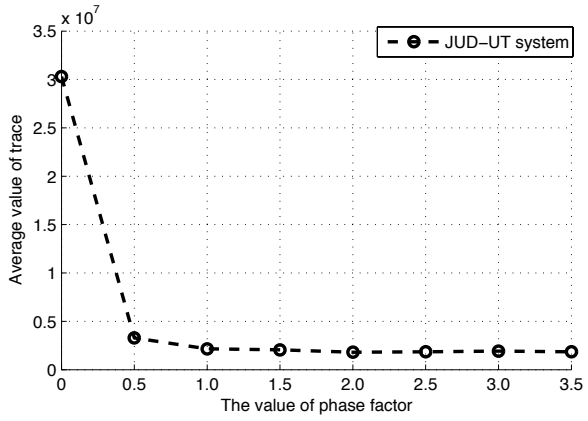

Figure 2: Average trace of covariance matrix $\mathbf{\Sigma}_i + \mathbf{\Sigma}_b^{(i)}$ respect to the phase term $\boldsymbol{\alpha}$ for JUD with UT compensation for SGMM systems.

the covariance of JUD-VTS system is very close to that of JUD-UT system, and the plot is omitted in the figure for clarity. The average value of covariance shows a similar trend to that of the WER when using different values of phase factors. This is not unexpected if one interprets the value of covariance as indicating the degree of "uncertainty" of the model. A small covariance may indicate that the model is more confident to explain the data, and if this confidence is gained from more accurate model compensation, it is expected to result in lower WER. However, the absolute value in the figure is not intuitive as the features were first transformed into another feature space by the JUD transformation $(\mathbf{A}^{(i)}, \mathbf{b}^{(i)})$.

To investigate the effect of phase factors to different noise conditions, we show the results of JUD with UT and VTS system on 14 testing sets in Table 2 and 3. We observe similar trend between the two. For the clean test set A, introducing a non-zero phase term does not improve accuracy just as expected. However, for the test set C which is also clean but recorded using a desk-mounted microphone, a large value of phase term increases the accuracy significantly. Since the training data is recorded using a close-talking microphone, the mismatch between training and testing data is mainly the channel noise including reverberation, which is correlated with speech. This is consistent with the assumption that phase factors model the correlations between noise and speech, and may explain the gains here. Comparing the results of sets B and D (in which 6 different types of noise were added to the clean sets A and C), the optimal values of the phase term are a little larger for D, which is probably because there is more channel noise in D which requires larger phase terms to capture the correlations. We could gain further insight by an analysis of the effect of the phase factor in different signal-to-noise-ratio (SNR) conditions, but unfortunately we cannot do this using the Aurora 4 corpus.

5. Conclusion

In this paper, we applied the unscented transform (UT) to estimate the nonlinear mismatch function for noise robust speech recognition based on subspace Gaussian mixture models (SGMMs). To reduce computational cost, we employed UT in the framework of joint uncertainty decoding (JUD). In this framework, we compared the performance of UT to first-order vector Taylor series (VTS) approximation by using the same noise model or with its own noise model estimation. Based on the Aurora 4 dataset, we show that JUD-UT can lead to better estimation accuracy than JUD-VTS system. We also discussed the effect of phase factor for JUD-UT system, and observe the similar trend compared to JUD-VTS system. By tuning the value of phase factors empirically, we can obtain similar accuracy by the two systems. We also analysed the effect of phase factors in terms of WERs in different noise conditions, and model covariance. Future work may include noise adaptive training [17] with both VTS and UT noise model estimation and compensation.

6. References

[1] PJ Moreno, B Raj, and RM Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*. IEEE, 1996, vol. 2, pp. 733–736.

[2] A Acero, L Deng, T Kristjansson, and J Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000.

[3] MJF Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, Cambridge University, 1995.

[4] SJ Julier and JK Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[5] Y Hu and Q Huo, "An HMM compensation approach using unscented transformation for noisy speech recognition," *Chinese Spoken Language Processing*, pp. 346–357, 2006.

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

[7] H Xu and KK Chin, "Comparison of estimation techniques in joint uncertainty decoding for noise robust speech recognition," in *Proc. INTERSPEECH*, 2009, pp. 2403–2406.

[8] J Li, D Yu, Y Gong, and L Deng, "Unscented transform with online distortion estimation for HMM adaptation," in *Proc. INTERSPEECH*, 2010.

Table 2: WERs of each test set with regards to the value of phase factor for JUD with UT system. "restau." indicates restaurant noise condition.

	A	B						C	D					
α	clean	car	babble	restau.	street	airport	station	clean	car	babble	restau.	street	airport	station
0	**4.9**	7.6	14.0	20.4	15.5	12.5	16.0	18.4	15.6	27.8	31.7	30.2	26.8	29.3
0.5	5.0	7.5	13.2	18.8	14.4	11.9	15.4	14.8	14.1	24.8	30.1	27.6	25.1	27.6
1.0	4.9	**7.3**	12.7	18.0	14.4	12.1	**15.2**	12.8	12.7	24.3	29.1	26.9	24.4	26.8
1.5	5.1	7.4	**12.4**	**17.9**	14.4	**11.7**	15.4	11.7	12.5	23.5	29.3	26.7	23.3	**26.3**
2.0	5.0	7.4	12.4	18.1	**14.1**	11.7	15.6	11.0	12.7	23.4	**28.8**	26.3	22.9	26.5
2.5	5.4	7.4	12.5	18.1	14.1	11.7	15.5	10.6	**12.3**	**23.2**	28.8	26.0	22.8	26.4
3.0	5.4	7.4	12.9	18.4	14.2	11.9	16.0	**10.1**	12.4	23.2	28.8	**25.8**	22.8	26.9
3.5	5.6	7.4	13.1	18.3	14.4	12.0	16.0	10.1	12.4	23.2	29.0	26.0	**22.7**	27.2

Table 3: WERs of each test set with regards to the value of phase factor for JUD with VTS system.

	A	B						C	D					
α	clean	car	babble	restau.	street	airport	station	clean	car	babble	restau.	street	airport	station
0	5.1	7.6	14.5	22.0	15.9	12.7	17.1	19.8	19.0	29.8	34.4	33.5	29.4	32.0
0.5	**5.0**	7.3	13.5	19.2	14.7	11.8	15.3	17.5	15.1	26.5	31.1	29.4	26.2	28.9
1.0	5.1	7.4	13.0	18.0	14.4	11.7	**15.2**	15.2	13.8	24.5	30.0	27.5	24.7	27.3
1.5	5.1	**7.2**	12.7	18.2	**13.9**	**11.5**	15.2	13.6	13.0	23.6	29.3	26.8	24.0	26.6
2.0	5.0	7.4	12.5	**17.8**	14.4	11.6	15.5	12.5	13.1	23.7	29.1	26.6	23.5	**26.3**
2.5	5.0	7.4	**12.4**	17.9	14.1	11.6	15.6	11.7	12.9	**23.4**	29.1	**25.9**	23.1	26.5
3.0	5.1	7.4	12.7	18.0	14.1	11.7	15.9	11.2	**12.7**	23.4	**28.8**	26.0	23.1	26.9
3.5	5.2	7.4	12.9	18.2	14.2	11.7	16.1	**10.9**	12.8	23.6	29.2	26.2	**23.0**	27.4

[9] F Gustafsson and G Hendeby, "Some relations between extended and unscented Kalman filters," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 545–555, 2012.

[10] D Povey, L Burget, M Agarwal, P Akyazi, F Kai, A Ghoshal, O Glembek, N Goel, M Karafiát, A Rastrow, RC Rose, P Schwarz, and S Thomas, "The subspace Gaussian mixture model—A structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[11] H Liao and MJF Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. INTERSPEECH*. Citeseer, 2005.

[12] 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 Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 133–143, 2004.

[13] J Li, L Deng, D Yu, Y Gong, and A Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer Speech & Language*, vol. 23, no. 3, pp. 389–405, 2009.

[14] RA Gopinath, MJF Gales, PS Gopalakrishnan, S Balakrishnan-Aiyer, and MA Picheny, "Robust speech recognition in noise—Performance of the IBM continuous speech recogniser on the ARPA noise spoke task," in *Proc. ARPA Workshops Spoken Lang. Syst. Technol.*, 1995, pp. 127–130.

[15] L Lu, KK Chin, A Ghoshal, and S Renals, "Noise compensation for subspace Gaussian mixture models," in *Proc. INTERSPEECH*, 2012.

[16] D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlıcek, Y Qian, P Schwarz, J Silovský, G Semmer, and K Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

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