HMM ADAPTATION USING A PHASE-SENSITIVE ACOUSTIC DISTORTION MODEL FOR ENVIRONMENT-ROBUST SPEECH RECOGNITION

Jinyu Li[1], Li Deng, Dong Yu, Yifan Gong, and Alex Acero

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052
jinyuli@ece.gatech.edu, {deng;dongyu;ygong;alexac}@microsoft.com

ABSTRACT

In this paper, we present a new approach to HMM adaptation that jointly compensates for additive and convolutive acoustic distortion in environment-robust speech recognition. The hallmark of our new approach is the use of a nonlinear, phase-sensitive model of acoustic distortion that captures phase asynchrony between clean speech and the mixing noise. In the first step of the developed algorithm, both the static and dynamic portions of the noise and channel parameters are estimated in the cepstral domain, using the speech recognizer's "feedback" information and the vector-Taylor-series linearization technique on the nonlinear phase-sensitive model. In the second step, the estimated noise and channel parameters are used to effectively adapt the static and dynamic portions of the HMM means and variances also using the linearized phase-sensitive acoustic distortion model.

In the experimental evaluation using the standard Aurora 2 task, the proposed new algorithm achieves 93.3% accuracy using the clean-trained complex HMM backend as the baseline system for unsupervised HMM adaptation. This reaches the highest performance number in the literature on this task with clean-trained HMM model. The experimental results show that the phase term, which was missing in all previous HMM-adaptation work, contributes significantly to the achieved high recognition accuracy.

Index Terms— phase-sensitive distortion model, vector Taylor series, additive and convolutive distortions, robust ASR

1. INTRODUCTION

Despite many years of research and investment, environment robustness in speech recognition remains an outstanding and difficult problem. In recent years, a popular approach to joint compensation of additive and convolutive distortions (JAC) in the model domain has been proposed and advanced [1][2][3][4][5], with promising results obtained. Common among these studies is the use of a parsimonious nonlinear "physical" model for environment distortion and the use of vector Taylor series (VTS) approximation to linearize or "Gaussianize" the model for closed-form HMM adaptation formulas and for noise/channel parameter estimation.

In all the previous JAC/VTS work for HMM adaptation, the environment-distortion model makes the simplifying assumption of instantaneous phase synchrony (phase-insensitive) between the clean speech and the mixing noise. This assumption was relaxed in the work reported in [6], where a new phase term was introduced to account for the random nature of the phase asynchrony. And it was shown in [6] that when the noise magnitude is estimated accurately, the Gaussian-distributed phase term plays a key role in recovering clean speech features by removing the noise and the cross term between the noise and speech.

However, in contrast to the JAC/VTS approach that implements robustness in the model (HMM) domain, the approach of [6] was implemented in the feature domain (i.e., feature enhancement instead of HMM adaptation), producing inferior recognition results than the model-domain approach despite the use of a more accurate environment-distortion model (phase-sensitive versus phase-insensitive models).

The research presented in this paper extends and integrates our earlier two sets of work: HMM adaptation with the phase-insensitive environment-distortion model (JAC/VTS [4][5]) and feature enhancement with the phase-sensitive environment-distortion model [6]. The new algorithm developed and presented in this paper implements environment robustness via HMM adaptation taking into account phase asynchrony between clean speech and the mixing noise. That is, it incorporates the same phase term in [6] into the rigorous formulation of JAC/VTS of [5]. We hence name our new algorithm as Phase-JAC/VTS.

The rest of the paper is organized as follows. In Section 2, we present the new Phase-JAC/VTS algorithm and its implementation steps. Experimental evaluation of the algorithm is provided in Section 3. We show that the new algorithm can achieve remarkably high (>93%) recognition accuracy averaged over all distortion conditions on the Aurora 2 task with the standard complex back-end, clean-trained model and standard MFCCs. We summarize our study and draw conclusions in Section 4.

2. PHASE-JAC/VTS ADAPTATION ALGORITHM

In this section, we first derive the Phase-JAC/VTS formulas for the HMM means and variances in the MFCC (both static and dynamic) domain using VTS approximation assuming that the estimates of the additive and convolutive parameters are known. We then give the algorithm which jointly estimates the additive and convolutive distortion parameters based on the same VTS approximation. A summary description follows on the detailed implementation steps of the entire algorithm which were used in our experiments.

2.1 Algorithm for HMM Adaptation Given the Joint Noise and Channel Estimates

Figure 1 shows a model for degraded speech with both noise

[1] This work was carried out at Microsoft Research, Redmond while the first author worked as a student intern.

(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

$$y[m] = x[m]*h[m] + n[m]. \quad (1)$$

With discrete Fourier transformation (DFT), the following equivalent relations can be established in the frequency domain:

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

where *k* is the frequency-bin index in DFT given a fixed-length time window.

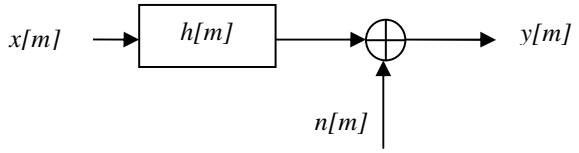

Figure 1: A model for acoustic environment distortion

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

$$|Y[k]|^2 = |X[k]|^2\,|H[k]|^2 + |N[k]|^2 + 2|X[k]|\,|H[k]||N[k]|\cos\theta_k \quad (3)$$

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

By applying a set of Mel-scale filters (*L* in total) to the power spectrum in Eq. (3), we have the *l*-th Mel filter-bank energies for distorted speech, clean speech, noise and channel:

$$|\tilde{Y}^{(l)}|^2 = \sum_k W_k^{(l)} |Y[k]|^2 \quad (4)$$

$$|\tilde{X}^{(l)}|^2 = \sum_k W_k^{(l)} |X[k]|^2 \quad (5)$$

$$|\tilde{N}^{(l)}|^2 = \sum_k W_k^{(l)} |N[k]|^2 \quad (6)$$

$$|\tilde{H}^{(l)}|^2 = \frac{\sum_k W_k^{(l)}\,|X[k]|^2\,|H[k]|^2}{|\tilde{X}^{(l)}|^2} \quad (7)$$

where the *l*-th filter is characterized by the transfer function $W_k^{(l)} \ge 0$ ($\sum_k W_k^{(l)} = 1$).

The phase factor $\alpha^{(l)}$ of the *l*-th Mel filter-bank is [6]:

$$\alpha^{(l)} = \frac{\sum_k W_k^{(l)}\,|X[k]|\,|H[k]|\,|N[k]|\cos\theta_k}{|\tilde{X}^{(l)}|\,|\tilde{H}^{(l)}|\,|\tilde{N}^{(l)}|}. \quad (8)$$

Then, the following relation is obtained in the Mel filter-bank domain for the *l*-th Mel filter-bank output [6]:

$$|\tilde{Y}^{(l)}|^2 = |\tilde{X}^{(l)}|^2\,|\tilde{H}^{(l)}|^2 + |\tilde{N}^{(l)}|^2 + 2\tilde{\alpha}^{(l)}|\tilde{X}^{(l)}|\,|\tilde{H}^{(l)}|\,|\tilde{N}^{(l)}|. \quad (9)$$

The phase-factor vector for all the *L* Mel filter-banks is defined as: $\boldsymbol{\alpha} = \left[\alpha^{(1)}, \alpha^{(2)}, ..., \alpha^{(l)}, ...\alpha^{(L)}\right]^T$. (10)

By taking logarithm and multiplying the non-square discrete cosine transform (DCT) matrix *C* to both sides of Eq. (9) for all the *L* Mel filter-banks, the following nonlinear distortion model is obtained in the cepstral domain:

$$y = x + h + C\log(1 + \exp(C^{-1}(n-x-h)) + 2\boldsymbol{\alpha}\bullet\exp(C^{-1}(n-x-h)/2))$$
$$= x + h + g_{\boldsymbol{\alpha}}(x,h,n), \quad \text{where} \quad (11)$$

$$g_{\boldsymbol{\alpha}}(x,h,n) = C\log(1 + \exp(C^{-1}(n-x-h)) + 2\boldsymbol{\alpha}\bullet\exp(C^{-1}(n-x-h)/2)) \quad (12)$$

and C^{-1} is the (pseudo) inverse DCT matrix. *y*, *x*, *n* and *h* are the vector-valued distorted speech, clean speech, noise, and channel, respectively, all in the MFCC domain. The $\bullet$ operation for two vectors denotes element-wise product, and each exponentiation of a vector above is also an element-wise operation.

Using the first-order VTS approximation with respect to *x*, *n* and *h*, and assuming the phase-factor vector $\boldsymbol{\alpha}$ is independent of *x*, *n* and *h*, we have

$$y \approx \mu_x + \mu_h + g(\mu_x,\mu_h,\mu_n) + G(x-\mu_x) + G(h-\mu_h) + (I-G)(n-\mu_n) \quad (13)$$

where $\frac{\partial y}{\partial x}\Big|_{\mu_x,\mu_n,\mu_h} = \frac{\partial y}{\partial h}\Big|_{\mu_x,\mu_n,\mu_h} = G$, (14)

$$\frac{\partial y}{\partial n} = I - G, \quad (15)$$

$$G = I - C\ diag\left(\frac{\exp(C^{-1}(\mu_n-\mu_x-\mu_h)) + \boldsymbol{\alpha}\bullet\exp(C^{-1}(\mu_n-\mu_x-\mu_h)/2)}{1+\exp(C^{-1}(\mu_n-\mu_x-\mu_h)) + 2\boldsymbol{\alpha}\bullet\exp(C^{-1}(\mu_n-\mu_x-\mu_h)/2)}\right)C^{-1} \quad (16)$$

and *diag(.)* stands for the diagonal matrix with its diagonal component value equal to the value of the vector in the argument. Each division of a vector is also an element-wise operation.

For the given noise mean vector μ_n and channel mean vector μ_h, the value of *G(.)* depends on mean vector μ_x. Specifically, for the *k*-th Gaussian in the *j*-th state, the element of *G(.)* matrix becomes:

$$G_{\boldsymbol{\alpha}}(j,k) = I - C\cdot diag\left(\frac{\exp(C^{-1}(\mu_n-\mu_{x,jk}-\mu_h)) + \boldsymbol{\alpha}\bullet\exp(C^{-1}(\mu_n-\mu_{x,jk}-\mu_h)/2)}{1+\exp(C^{-1}(\mu_n-\mu_{x,jk}-\mu_h)) + 2\boldsymbol{\alpha}\bullet\exp(C^{-1}(\mu_n-\mu_{x,jk}-\mu_h)/2)}\right)\cdot C^{-1} \quad (17)$$

Then, the Gaussian mean vectors (the *k*-th Gaussian in the *j*-th state) in the adapted HMM for the degraded speech can be obtained by taking expectation of both sides of Eq. (13):

$$\mu_{y,jk,\boldsymbol{\alpha}} \approx \mu_{x,jk} + \mu_h + g_{\boldsymbol{\alpha}}(\mu_{x,jk},\mu_h,\mu_n), \quad (18)$$

which is applied only to the static portion of the MFCC vector.

The covariance matrix $\Sigma_{y,jk,\boldsymbol{\alpha}}$ in the adapted HMM can be estimated as a weighted sum of $\Sigma_{x,jk}$, the covariance matrix of the clean HMM, and Σ_n, the covariance matrix of noise, by taking variance "operation" on both sides of Eq. (13):

$$\Sigma_{y,jk,\boldsymbol{\alpha}} \approx G_{\boldsymbol{\alpha}}(j,k)\Sigma_{x,jk}G_{\boldsymbol{\alpha}}(j,k)^T + (I-G_{\boldsymbol{\alpha}}(j,k))\Sigma_n(I-G_{\boldsymbol{\alpha}}(j,k))^T \quad (19)$$

Here, no channel variance is taken into account because we treat the channel as a fixed, deterministic quantity in a given utterance.

For the delta and delta/delta portions of MFCC vectors, the adaptation formulas for the mean vector and covariance matrix are

$$\mu_{\Delta y,jk,\boldsymbol{\alpha}} \approx G_{\boldsymbol{\alpha}}(j,k)\mu_{\Delta x,jk} + (I-G_{\boldsymbol{\alpha}}(j,k))\mu_{\Delta n}, \quad (20)$$

$$\mu_{\Delta\Delta y,jk,\boldsymbol{\alpha}} \approx G_{\boldsymbol{\alpha}}(j,k)\mu_{\Delta\Delta x,jk} + (I-G_{\boldsymbol{\alpha}}(j,k))\mu_{\Delta\Delta n}, \quad (21)$$

$$\Sigma_{\Delta y,jk,\boldsymbol{\alpha}} \approx G_{\boldsymbol{\alpha}}(j,k)\Sigma_{\Delta x,jk}G_{\boldsymbol{\alpha}}(j,k)^T + (I-G_{\boldsymbol{\alpha}}(j,k))\Sigma_{\Delta n}(I-G_{\boldsymbol{\alpha}}(j,k))^T, \quad (22)$$

$$\Sigma_{\Delta\Delta y,jk,\boldsymbol{\alpha}} \approx G_{\boldsymbol{\alpha}}(j,k)\Sigma_{\Delta\Delta x,jk}G_{\boldsymbol{\alpha}}(j,k)^T + (I-G_{\boldsymbol{\alpha}}(j,k))\Sigma_{\Delta\Delta n}(I-G_{\boldsymbol{\alpha}}(j,k))^T. \quad (23)$$

2.2 Algorithm for Re-estimation of Noise and Channel

EM algorithm is developed as part of the overall Phase-JAC/VTS algorithm to estimate all the noise and channel parameters using the first order VTS approximation. Let $\gamma_t(j,k)$ denote the posterior probability for the *k*-th Gaussian in the *j*-th state of the HMM, i.e., $\gamma_t(j,k) = p(\theta_t = j, \varepsilon_t = k | Y, \bar{\lambda})$, (24)

where θ_t denote the state index, and ε_t denote the Gaussian index at time frame *t*. $\bar{\lambda}$ is the old parameter sets of noise and channel. The re-estimation formulas for the static channel mean μ_h, the static and dynamic noise means [μ_n, $\mu_{\Delta n}$, $\mu_{\Delta\Delta n}$], and the static and dynamic noise variances [Σ_n, $\Sigma_{\Delta n}$, $\Sigma_{\Delta\Delta n}$] are (derivations omitted):

$$\mu_h = \mu_{h,0} + \left\{\sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) G_\alpha(j,k)^T \Sigma_{y,jk,\alpha}^{-1} G_\alpha(j,k)\right\}^{-1} \cdot \left\{\sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) G_\alpha(j,k)^T \Sigma_{y,jk,\alpha}^{-1} \left[y_t - \mu_{x,jk} - \mu_{h,0} - g_\alpha(\mu_{x,jk}, \mu_{h,0}, \mu_{n,0})\right]\right\} \quad (25)$$

$$\mu_n = \mu_{n,0} + \left\{\sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{y,jk,\alpha}^{-1} (I - G_\alpha(j,k))\right\}^{-1} \cdot \left\{\sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{y,jk,\alpha}^{-1} \left[y_t - \mu_{x,jk} - \mu_{h,0} - g_\alpha(\mu_{x,jk}, \mu_{h,0}, \mu_{n,0})\right]\right\} \quad (26)$$

$$\mu_{\Delta n} = \mu_{\Delta n,0} + \left\{\sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{\Delta y,jk,\alpha}^{-1} (I - G_\alpha(j,k))\right\}^{-1} \cdot \left\{\sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{\Delta y,jk,\alpha}^{-1} \left[\Delta y_t - G_\alpha \mu_{\Delta x,jk} - (I - G_\alpha(j,k)) \mu_{\Delta n,0}\right]\right\} \quad (27)$$

$$\mu_{\Delta\Delta n} = \mu_{\Delta n,0} + \left\{\sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{\Delta\Delta y,jk,\alpha}^{-1} (I - G_\alpha(j,k))\right\}^{-1} \cdot \left\{\sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{\Delta\Delta y,jk,\alpha}^{-1} \left[\Delta\Delta y_t - G_\alpha \mu_{\Delta\Delta x,jk} - (I - G_\alpha(j,k)) \mu_{\Delta\Delta n,0}\right]\right\} \quad (28)$$

$$\Sigma_n = \Sigma_{n,0} - \left(\frac{\partial^2 Q}{\partial^2 \Sigma_n}\right)_{\Sigma_n=\Sigma_{n,0}}^{-1} \left(\frac{\partial Q}{\partial \Sigma_n}\right)_{\Sigma_n=\Sigma_{n,0}} . \quad (29)$$

$\Sigma_{\Delta n}$ and $\Sigma_{\Delta\Delta n}$ are updated in a similar way to Eq. (29) by replacing the static parameters with the corresponding delta and delta/delta parameters (derivations omitted here).

2.3 Algorithm implementation

The implementation steps for the JAC/VTS HMM adaptation algorithm described so far in this section and used in our experiments are summarized and described in the following:

1. Read in a distorted speech utterance;
2. Set the channel mean vector to all zeros;
3. Initialize the noise mean vector and diagonal covariance matrix using the first and last *N* frames (speech-free) from the utterance using sample estimates;
4. Compute the Gaussian- and $\boldsymbol{\alpha}$-dependent *G(.)* with (17), and update/adapt the HMM parameters with (18)–(23);
5. Decode the utterance with the adapted HMM parameters;
6. Compute posterior probabilities of (24) and then re-estimate all the noise and channel parameters with (25)-(29);
7. Compute *G(.)* with (17), and update/adapt the HMM parameters with (18)–(23);
8. Use the final adapted model to obtain the utterance output transcription;
9. Goto step 1.

A challenging problem in Phase-JAC/VTS is the setting of the phase-factor vector, $\boldsymbol{\alpha}$. In Section 2.1, we assumed $\boldsymbol{\alpha}$ is independent of speech, noise, and channel (This assumption will be removed in our future study). And in current implementation, each component of $\boldsymbol{\alpha}$ is also assumed to be a fixed, tunable value, α, i.e., $\tilde{\alpha}^{(l)} = \alpha$. In the experiment section, varying values of α are chosen to evaluate Phase-JAC/VTS. Our earlier JAC/VTS [5] can be considered as a special case of the current Phase-JAC/VTS when $\alpha = 0$ uniformly.

The steps above are for one pass decoding and one-iteration EM re-estimation of noise/channel parameters, as we have carried out so far in our experiments to be presented in the next section. For multiple-pass decoding (as will be reported in future publications), there would be a loop between steps 5 and 7, and multiple-iteration EM for noise/channel re-estimation would be implemented by looping between steps 6 and 7.

3. SPEECH RECOGNITION EXPERIMENTS

The effectiveness of the Phase-JAC/VTS algorithm presented in Section 2 has been evaluated on the standard Aurora 2 task [7] 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. The data in set-A and set-B contain eight different types of additive noise, while set-C contain two different types of noise plus additional channel distortion. The baseline experiment setup follows the standard script provided by ETSI, including the standard complex "backend" [8] 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 0 is used instead of the log energy in the original script.

The Phase-JAC/VTS 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 detailed implementation steps described in Section 2.3 are used in the experiments. We use the first and last *N*=20 frames from each utterance for initializing the noise means and variances. Only one pass processing is used in the reported experiments.

The theory developed in [6] has shown that given true noise and channel parameters, the range of α value is between -1 and 1 in theory. To take into account inaccuracy in the noise/channel estimates, we widened the range of the α value, which was set up to 5 (with an interval of 0.25). The corresponding recognition accuracies (Accs) are plotted in Figure 2. The results are somewhat surprising in two ways. First, the optimal value is $\alpha = 2.5$, significantly beyond the normal range between -1 and 1 (see detailed discussions below). Second, the recognition accuracy at $\alpha = 2.5$, 93.32%, is much higher than the use of phase-insensitive distortion model for JAC/VTS (equivalent to setting $\alpha = 0$ in Figure 2), demonstrating the critical role of the use of phase asynchrony between clean speech and the mixing noise. Table 1 lists detailed test results for clean-trained complex backend HMM system after Phase-JAC/VTS adaptation with the optimal α value.

The optimal performance achieved at $\alpha = 2.5$ seems to have contradicted the theory in [6] that α should be less than 1. We offer two possible reasons here. First, the theory in [6] is built on the basis that the correct noise and channel vectors are given. For

Phase-JAC/VTS, the noise and channel are estimated with possibly systematic biases, because the truncated VTS discards the second and all higher-order terms. A larger α may be used partly to compensate for these biased estimates. (More detailed analyses on this are provided in [9]). Second, by definition of (8), α is a random variable, due to the random speech/noise mixing phase θ_k, instead of a deterministic one as used in this study. Extending the current work by including variance of α may move the optimal range of α values back closer to the normal, expected range of lower than one.

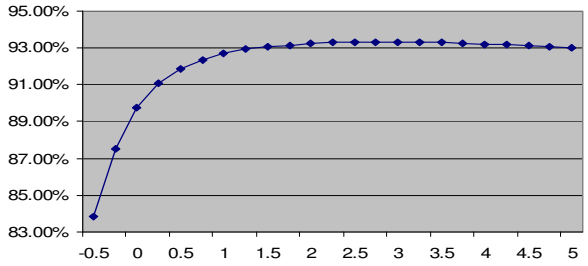

Figure 2: Aurora 2 recognition accuracy for the Phase-JAC/VTS algorithm as a function of the α value.

Table 1: Detailed Aurora 2 accuracy of clean-trained complex backend HMMs after adaptation using Phase-JAC/VTS where $\alpha = 2.5$.

Clean Training - Results														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
20 dB	99.14	99.03	99.52	99.11	99.20	99.26	99.03	99.58	99.51	99.34	99.39	99.06	99.22	**99.26**
15 dB	98.99	98.55	99.05	98.86	98.86	98.77	98.67	99.25	99.07	98.94	98.86	98.58	98.72	**98.84**
10 dB	97.57	96.86	97.76	96.36	97.14	96.28	97.31	97.7	97.84	97.28	97.64	97.1	97.37	**97.26**
5 dB	94.23	90.75	94.84	92.41	93.06	90.85	92.38	93.89	93.34	92.62	94.6	92.11	93.35	**93.01**
0 dB	80.56	68.74	83.48	80.16	78.24	70.95	78.87	80.7	79.98	77.62	82.87	76.75	79.81	**78.56**
Average	94.10	90.79	94.93	93.38	**93.30**	91.22	93.25	94.22	93.95	**93.16**	94.67	92.72	**93.70**	**93.32**

4. CONCLUSION

In this paper, we have presented our recent development of the Phase-JAC/VTS algorithm for HMM adaptation and demonstrated its effectiveness in the standard Aurora 2 environment-robust speech recognition task. The algorithm distinguishes itself from all previous related work by introducing the novel phase term in JAC model of environmental distortion for on HMM adaptation. We derived the estimation formulas for all noise and channel parameters and the adaptation formulas for all static and dynamic HMM parameters in the same framework of Phase-JAC/VTS.

In the experimental evaluation using the standard Aurora 2 task, the proposed Phase-JAC/VTS algorithm has achieved 93.32% accuracy using the clean-trained complex HMM backend as the baseline system for the model adaptation. This reaches the highest performance number in the literature on this task without discriminative training of the HMM system. The experimental results have shown that the value of the phase-factor vector is critical to the success of Phase-JAC/VTS.

Several research issues need to be addressed in the future to further increase the effectiveness of the algorithm presented in this paper. First, the α value is chosen manually and is set as same for all utterances in this study. An utterance-dependent strategy for setting α should be derived. Second, the phase-factor vector, $\boldsymbol{\alpha}$, is set to have a constant α value for its every component. By examining Eqs. (8) and (10), it is easy to see components of $\boldsymbol{\alpha}$ have different values. Third, as shown in [6], instead of being a constant vector, $\boldsymbol{\alpha}$ should follow a distribution and this will change the current algorithm in a significant way. Fourth, as analyzed in the experiment section, biased estimates of noise and channel may result in the unusual optimal values of α. We need to examine whether the α value fits the theoretically range as analyzed in [6] after obtaining more reliable estimates of noise and channel. Resolving the above issues, we expect to achieve greater effectiveness of the Phase-JAC/VTS algorithm than what has been reported in this paper.

5. ACKNOWLEDGEMENTS

We would like to thank Dr. Jasha Droppo at Microsoft research for the help in setting up the experimental platform.

6. REFERENCES

[1] Y. Gong, "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.

[2] 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.

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

[4] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," *Proc. ICSLP*, Vol.3, pp. 869-872, 2000.

[5] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions," *Proc. IEEE ASRU*, Dec. 2007, accepted and to appear.

[6] L. Deng, J. Droppo, and A. Acero. "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.

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

[8] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouver, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," *Proc. ICSLP*, pp. 17–20, 2002.

[9] L. Deng (invited), "Roles of high-fidelity acoustic modeling in robust speech recognition," paper invited to present at the IEEE ASRU Workshop, Dec. 2007, pp. 1-12.