Feature Space Generalized Variable Parameter HMMs for Noise Robust Recognition

Yang Li[1], *Xunying Liu*[2] *& Lan Wang*[1]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences /The Chinese University of Hong Kong, Hong Kong, China
[2]Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.

Abstract

Handling variable ambient noise is a challenging task for automatic speech recognition (ASR) systems. To address this issue, multi-style training using speech data collected in diverse noise environments, noise adaptive training or uncertainty decoding techniques can be used. An alternative approach is to explicitly approximate the continuous trajectory of Gaussian component or model space linear transform parameters against the varying noise, for example, using generalized variable parameter HMMs (GVP-HMM). In order to reduce the computational cost of conventional GVP-HMMs when model parameter update against the varying noise condition is required, this paper investigates a novel and more efficient extension of GVP-HMMs that can also model the trajectories of feature space linear transforms. Significant error rate reductions of 9.3% and 18.5% relative were obtained over the multi-style training baseline system on Aurora 2 and a medium vocabulary Mandarin Chinese speech recognition task respectively.

Index Terms: feature transform trajectory, generalized variable parameter HMM, variable noise, robust speech recognition

1. Introduction

The presence of environmental noise often leads to severe degradation of automatic speech recognition (ASR) performance. When the noise condition is of time varying nature, this problem becomes even more challenging. To handle this issue, a range of model based techniques can be used: multi-style training uses speech data collected in a wide range of diverse noise environments [16], and exploits the implicit modelling ability of mixture models, and more recently deep neural networks [19], to obtain a good generalization to unseen noise conditions; noise adaptive training [11, 8, 12] structurally models the variability introduced to the observed speech signals by environment noise and other factors; uncertainty decoding (UD) [6, 17, 7, 14, 20, 21], rather than using a point estimate of the corrupted features, propagate the uncertainty that varies with the noise represented by, for example, a conditional distribution of the corrupted speech, into the recognizer. In addition to the above approaches, it is also possible to explicitly approximate the continuous trajectories of optimal model parameters against the varying noise condition using a polynomial function [9, 4, 25, 15], for example, as in multiple regression HMMs (MR-HMM) [9] and variable parameter HMMs (VP-HMM) [4, 23, 24, 25].

This work is supported by National Natural Science Foundation of China (NSFC 61135003, NSFC 90920002), and Guangdong Innovative Research Team Program (No. 201001D0104648280).

In order to reduce the interpolation cost incurred at Gaussian component level when mean or variance trajectory modelling are used, an extension to both MR-HMMs and VP-HMMs, the generalized variable parameter HMM (GVP-HMM), was proposed in [2, 3]. In addition to Gaussian means and variances, GVP-HMMs can also provide a more compact trajectory modelling for model space tied linear transformations, and thus provide a flexible form of parameter trajectory modelling. For example, when only limited amounts of noisy training data is available, to ensure all polynomial coefficients are robustly estimated, only the trajectories associated with the elements of a globally tied mean transform can be considered. When large amounts of noisy training data is used, a more refined modelling resolution can also be obtained by increasing the number of tied transformations, or modelling the trajectories of multiple parameter types simultaneously. However, as the underlying noise condition varies frequently against time, applying the resulting updated linear transforms to all Gaussian mean parameters becomes highly expensive.

To address this issue, this paper investigates a novel and more efficient extension of GVP-HMMs that can also model the trajectories of feature space linear transform parameters against the varying noise. As the updated transforms are directly applied to the acoustic features, rather than Gaussian component means, the computational cost when using GVP-HMMs are thus expected to be significantly reduced. The rest of the paper is organized as follows. The GVP-HMM framework is reviewed in section 2. Feature space GVP-HMMs are proposed in section 3. A range of GVP-HMM systems using various modelling configurations are described in section 4. In section 5 various model and feature space GVP-HMM based noise compensation schemes are evaluated on Aurora 2 and a medium vocabulary Mandarin Chinese speech recognition task. Section 6 is the conclusion and future research.

2. Generalized Variable Parameter HMMs

Generalized variable parameter HMMs (GVP-HMMs) [2, 3] explicitly model the trajectory of optimal acoustic parameters that vary with respect to the underlying noise condition. The type of parameter trajectories are not restricted to those of means and covariances of conventional tied mixture HMMs. Other more compact forms of parameters, such as model or feature space linear transformations [13, 10], may also be considered. In previous research, only trajectories of Gaussian mean transforms were modelled [2, 3]. For a D dimensional observation $\boldsymbol{o}_t$ emitted from Gaussian mixture component m, assuming

P^{th} order polynomials are used, this is given by

$$\boldsymbol{o}^{(t)} \sim p\left(\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \boldsymbol{\Sigma}^{(m)}(\mathbf{v}_t), \boldsymbol{W}^{(r_m)}(\mathbf{v}_t)\right) \quad (1)$$

where $\mathbf{v}_t^\top$ is a $(P+1)$ dimensional Vandermonde vector [1], such that $\mathbf{v}_{t,p} = v_t^{p-1}$. v_t is an auxiliary feature, and in this paper, the speech-noise-ratio (SNR) condition [18] at frame t. $\boldsymbol{W}^{(r_m)}(\mathbf{v}_t)$ is the $(D+1)\times D$ mean transform that component m is assigned to at frame t. $\boldsymbol{\mu}^{(m)}(\cdot)$, $\boldsymbol{\Sigma}^{(m)}(\cdot)$ and $\boldsymbol{W}^{(r_m)}(\cdot)$ are the P^{th} order mean, covariance and MLLR mean transform trajectory polynomials of component m respectively. Assuming diagonal covariances are used, then the trajectories of the i^{th} dimension of the mean and variance, and the transform element in row i and column j, are

$$\begin{aligned} \mu_i^{(m)}(\mathbf{v}_t) &= \mathbf{v}_t \cdot \mathbf{c}^{(\mu_i^{(m)})} \\ \sigma_{i,i}^{(m)}(\mathbf{v}_t) &= \breve{\sigma}_{i,i}^{(m)} \mathbf{v}_t \cdot \mathbf{c}^{(\sigma_{i,i}^{(m)})} \\ w_{i,j}^{(r_m)}(\mathbf{v}_t) &= \mathbf{v}_t \cdot \mathbf{c}^{(w_{i,j}^{(r_m)})} \end{aligned} \quad (2)$$

where $\mathbf{c}^{(\cdot)}$ is a $(P+1)$ dimensional polynomial coefficient vector such that $\mathbf{c}_p^{(\cdot)} = c_{p-1}^{(\cdot)}$, and $c_{p-1}^{(\cdot)}$ the $(p-1)^{th}$ order polynomial coefficient of the parameter trajectory being considered. $\breve{\sigma}_{i,i}^{(m)}$ is the clean speech based variance estimate. By definition, the mean transform polynomials are modelled on top of the component mean trajectories, thus the final updated mean vector of component m at time instance t is computed as

$$\hat{\boldsymbol{\mu}}^{(m)}(\mathbf{v}_t) = \boldsymbol{W}^{(r_m)}(\mathbf{v}_t)\boldsymbol{\zeta}_t^{(m)} \quad (3)$$

where the $(D+1)$ dimensional extended mean vector trajectory $\boldsymbol{\zeta}_t^{(m)} = [\boldsymbol{\mu}^{(m)}(\mathbf{v}_t), 1]^\top$.

GVP-HMMs share the same instantaneous adaptation power as standard MR-HMMs and VP-HMMs. For any noise condition, present or unseen in the training data, GVP-HMMs can instantly produce the matching Gaussian component and mean transform parameters by-design without requiring any multi-pass decoding and adaptation process.

3. Feature Space GVP-HMMs

As discussed in section 1, when using mean transform based GVP-HMMs, as the underlying noise condition varies frequently in time, for example, at segment or frame level, applying the resulting updated linear transforms to Gaussian component means, which often can be in a large number for complex systems, become highly expensive. The solution considered in this paper is to extend the GVP-HMM framework presented in section 2 to also model the trajectories of feature space linear transform parameters against the varying noise condition. Applying the resulting updated linear transforms to the acoustic features, rather than component means, is expected to significantly reduce the computational cost in recognition. Hence, the form of GVP-HMMs in equation (1) is modified as

$$\boldsymbol{o}^{(t)} \sim p\left(\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \boldsymbol{\Sigma}^{(m)}(\mathbf{v}_t), \boldsymbol{W}^{(r_m)}(\mathbf{v}_t), \widetilde{\boldsymbol{W}}^{(r_m)}(\mathbf{v}_t)\right) \quad (4)$$

where $\widetilde{\boldsymbol{W}}^{(r_m)}(\mathbf{v}_t)$ is a $(D+1)\times D$ feature space transform that component m is assigned to at frame t. The polynomial trajectory of the transform element in row i and column j is

$$\widetilde{w}_{i,j}^{(r_m)}(\mathbf{v}_t) = \mathbf{v}_t \cdot \mathbf{c}^{(\widetilde{w}_{i,j}^{(r_m)})} \quad (5)$$

The updated observation vector for component m at time instance t is computed as

$$\hat{\boldsymbol{o}}^{(t,r_m)} = \widetilde{\boldsymbol{W}}^{(r_m)}(\mathbf{v}_t)\widetilde{\boldsymbol{\zeta}}_t \quad (6)$$

where the $(D+1)$ dimensional extended observation vector $\widetilde{\boldsymbol{\zeta}}_t = [\boldsymbol{o}^{(t)}, 1]^\top$.

In common with conventional MR-HMMs, VP-HMMs and GVP-HMMs, the feature transform polynomial coefficients $\mathbf{c}^{(\widetilde{w}_{i,j}^{(r_m)})}$ can also be estimated using multi-style training [16, 2, 3] on diverse speech data that covers a range of observed noise conditions. The associated maximum likelihood (ML) auxiliary function is given by [5],

$$\begin{aligned} \mathcal{Q}(\lambda, \tilde{\lambda}) = \sum_{m,t} \gamma_m(t) \log p\Big(\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \\ \boldsymbol{\Sigma}^{(m)}(\mathbf{v}_t), \boldsymbol{W}^{(r_m)}(\mathbf{v}_t), \widetilde{\boldsymbol{W}}^{(r_m)}(\mathbf{v}_t)\Big) \end{aligned} \quad (7)$$

where $\gamma_m(t)$ is the posterior probability of frame $\boldsymbol{o}_t$ at component m.

Combining the above with equations (4), (5) and (6), setting the gradient against the polynomial coefficient vectors associated with the feature transform elements to zero, and also assuming there are a total of Q observed noise conditions, $\{v_1, ..., v_q, ..., v_Q\}$, found in the training data, the following row-by-row update formula can be derived,

$$\mathbf{c}^{(\widetilde{w}_i^{(r_m)})} = \mathbf{U}^{(\widetilde{w}_i^{(m)})-1}\left(\mathbf{k}^{(\widetilde{w}_i^{(m)})} + \sum_{m\in r_m, t}^{v_t=v_q} \alpha_q \gamma_m(t) \boldsymbol{\eta}_{q,i}^\top\right) \quad (8)$$

where $\mathbf{c}^{(\widetilde{w}_i^{(r_m)})}$ is a $(D+1)\times(P+1)$ dimensional meta polynomial coefficient vector spanning across all elements of row i of transform $\widetilde{\boldsymbol{W}}^{(r_m)}(\cdot)$, and the sufficient statistics $\mathbf{U}^{(\widetilde{w}_i^{(r_m)})}$ is a $[(D+1)\times(P+1)]\times[(D+1)\times(P+1)]$ meta Vandermonde matrix, and $\mathbf{k}^{(\widetilde{w}_i^{(r_m)})}$ a $(D+1)\times(P+1)$ dimensional meta regression target vector. These are computed as

$$\begin{aligned} \mathbf{U}^{(\widetilde{w}_i^{(m)})} &= \sum_{m\in r_m, t} \gamma_m(t) \sigma_{i,i}^{(m)-1}(\mathbf{v}_t) \boldsymbol{\theta}_{t,i}^\top \boldsymbol{\theta}_{t,i} \\ \mathbf{k}^{(\widetilde{w}_i^{(m)})} &= \sum_{m\in r_m, t} \gamma_m(t) \sigma_{i,i}^{(m)-1}(\mathbf{v}_t) \hat{\mu}_i^{(m)}(\mathbf{v}_t) \boldsymbol{\theta}_{t,i}^\top \end{aligned} \quad (9)$$

and the two $(D+1)\times(P+1)$ dimensional meta vectors,

$$\begin{aligned} \boldsymbol{\eta}_{t,i}^\top &= \left[cof\left(\widetilde{w}_{i,1}^{(m)}(\mathbf{v}_t)\right)\mathbf{v}_t, \ ..., cof\left(\widetilde{w}_{i,D+1}^{(m)}(\mathbf{v}_t)\right)\mathbf{v}_t\right]^\top \\ \boldsymbol{\theta}_{t,i}^\top &= \left[\widetilde{\boldsymbol{\zeta}}_{t,1}\mathbf{v}_t, \ ..., \widetilde{\boldsymbol{\zeta}}_{t,j}\mathbf{v}_t, \ ..., \widetilde{\boldsymbol{\zeta}}_{t,D+1}\mathbf{v}_t\right]^\top. \end{aligned} \quad (10)$$

where $cof(\widetilde{w}_{i,j}^{(r_m)}(\mathbf{v}_t))$ is the extended cofactor of feature space transform element $\widetilde{w}_{i,j}^{(r_m)}(\mathbf{v}_t)$.

For each observed noise condition q, the scaling factor α_q used in the update formula of equation (8) can be found by solving the quadratic equation below

$$\begin{aligned} \alpha_q^2\left(\sum_{m\in r_m, t, v_t=v_q} \gamma_m(t)\right)\left(\boldsymbol{\eta}_{q,i}\mathbf{U}_q^{(w_i^{(m)})-1}\boldsymbol{\eta}_{q,i}^\top\right) + \\ \alpha_q\left(\boldsymbol{\eta}_{q,i}\mathbf{U}_q^{(w_i^{(m)})-1}\mathbf{k}_q^{(w_i^{(m)})}\right) - 1 = 0. \end{aligned} \quad (11)$$

and choosing the solution that maximizes the auxiliary equation of equation (7), where the noise condition dependent sufficient statistics are accumulated as

$$\mathbf{U}_q^{(\tilde{w}_i^{(m)})} = \sum_{m \in r_m, v_t = v_q} \gamma_m(t) \sigma_{i,i}^{(m)-1}(\mathbf{v}_t) \boldsymbol{\theta}_{t,i}^{\top} \boldsymbol{\theta}_{t,i}$$

$$\mathbf{k}_q^{(\tilde{w}_i^{(m)})} = \sum_{m \in r_m, v_t = v_q} \gamma_m(t) \sigma_{i,i}^{(m)-1}(\mathbf{v}_t) \hat{\mu}_i^{(m)}(\mathbf{v}_t) \boldsymbol{\theta}_{t,i}^{\top}. \quad (12)$$

The update formula given in equation (8) are performed iteratively until the auxiliary function in equation (7) is converging.

4. Variants of GVP-HMM Systems

As discussed in sections 1 and 2, in order to adjust the trade-off between modelling resolution, robustness in estimation and computational efficiency, a wide rage of GVP-HMM configurations may be considered to suit different purposes. Description of these GVP-HMM variant systems' configurations and the number of polynomial coefficients used in a Mandarin Chinese acoustic model that contains 29k Gaussian components using 42 dimensional PLP and pitch features, are shown in table 1.

GVP System	Parameter Polynomials				#Poly Coef
	mean	var	tran	ftran	
mean	√	×	×	×	3.66M
mv	√	√	×	×	7.32M
tran2	×	×	√	×	10.8K
tran8	×	×	√	×	43.2k
tran256	×	×	√	×	1.39M
ftran2	×	×	×	√	10.8K
ftran8	×	×	×	√	43.2k
ftran256	×	×	×	√	1.39M
ftran8-tran8	×	×	√	√	86.4k
ftran256-tran256	×	×	√	√	2.79M
mv-tran2	√	√	√	×	7.32M
mv-ftran2	√	√	×	√	7.32M
mv-ftran2-tran2	√	√	√	√	7.32M

Table 1: Description of various GVP-HMMs: parameter polynomial types and the number of polynomial coefficients.

Two standard VP-HMM configurations, which allow trajectory modelling of Gaussian component means, and optionally variances, are shown in the first two lines of the table, as "mean" and "mv" respectively. In the 2nd section (line 3 to 5) of table 1, three GVP-HMM systems modelling the polynomial trajectories of 2, 8 or 256 mean transforms are shown as "tran2", "tran8" and "tran256". In the 3rd section (line 7 to 9) of the table, three comparable GVP-HMM systems modelling the polynomial trajectories of feature space transforms are shown as "ftran2", "ftran8" and "ftran256" respectively. Two GVP-HMMs systems that model the trajectories of both mean and feature transforms are shown as "ftran8-tran8" and "ftran256-tran256" in the 4th section (line 9 to 10) of table 1. Finally, three more complex GVP-HMMs systems that use trajectory modelling for Gaussian means and variances, plus 2 model or(and) feature space transforms are shown as "mv-tran2", "mv-ftran2" and "mv-ftran2-tran2" in the bottom section of the table.

5. Experimental Results

In this section, feature GVP-HMM systems are evaluated on two tasks: Aurora 2 and a medium vocabulary Mandarin Chinese In-car navigation command recognition task. All GVP-HMM system used second order polynomials for trajectory modelling in the experiments.

5.1. Experiments on Aurora 2

The Aurora2 speaker independent digit sequence recognition database contains 4 noisy conditions: subway, babble, car and exhibition. A total of 420 utterances from four different SNR conditions (-5dB, 5dB, 15dB, 25dB) were used to train both the baseline multi-style HMMs and various GVP-HMM systems. A total of 1000 utterances selected from the car noise environment at 0dB, 5dB, 10dB, 15dB and 20dB SNR were used for word error rate (WER) evaluation.

Performance of the multi-style baseline and various GVP-HMM systems, as described in table 1 are shown in table 2. Modelling both Gaussian mean and variance trajectories gave the best performance for standard VP-HMMs, as shown in the 3rd and 4th lines of table 2. Using the "mv" system, average WER reductions of 0.56%-0.58% absolute (6.3%-6.5% relative) across all SNR conditions were obtained over the "mcond" multi-style baseline, and the mean only VP-HMM/GVP-HMM system shown as "mean" in table 2.

System	0dB	5dB	10dB	15dB	20dB	Ave
clean baseline	75.33	41.42	15.63	6.14	3.14	28.34
mcond baseline	22.88	9.42	4.29	3.58	2.78	8.95
mean	25.63	9.52	4.32	3.10	2.26	8.97
mv	23.16	9.12	4.30	3.09	2.28	8.39
tran2	30.22	9.47	4.98	3.21	2.28	10.09
tran8	22.58	9.02	4.27	3.07	2.23	8.23
mv-tran2	22.34	8.96	4.18	3.04	2.29	8.16
ftran2	30.73	9.77	4.95	3.20	2.48	10.23
ftran8	22.75	9.05	4.28	3.10	2.46	8.33
mv-ftran2	22.84	9.01	4.20	3.07	2.32	8.29
ftran8-tran8	22.42	8.99	4.23	3.06	2.23	8.19
mv-ftran2-tran2	22.15	8.89	4.22	3.04	2.30	8.12

Table 2: WER performance of baseline and various GVP-HMM systems on Aurora 2

Feature space GVP-HMM systems were found to give error rates similar to the comparable model transform based GVP-HMM systems, as are shown in the 3rd and 4th sections of table 2. For example, the mean transform based GVP-HMM system "tran8", and the comparable feature space GVP-HMM system "ftran8", both outperformed the "mv" system and reduced the average WER to 8.23% and 8.33% respectively. The use of feature space GVP-HMMs were also found to incur lower computational when applying the updated linear transforms to the observations than mean transform based GVP-HMM systems, by 21.8% on average across different test SNR conditions, as are shown Fig 1 for the "tran8" and "ftran8" GVP-HMMs.

A more complex GVP-HMM system, "ftran8-tran8", that used both model and feature space transform trajectory modelling, further reduced error rate to 8.19%. When both Gaussian component and transform parameter trajectories are modelled, small further improvements were obtained. The "mv-ftran2-tran2" system outperformed the "mv" system by 0.23%-1.01% absolute on the 0dB and 5db data, and on average across all test SNR conditions, by 0.83% absolute (9.3% relative) over the multi-style trained baseline "mcond" model. It gave the lowest WER among all systems in table 2.

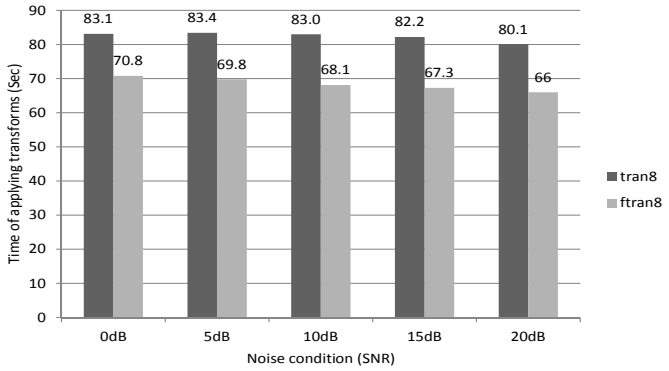

Figure 1: Computational cost of applying GVP-HMMs model-space transform/feature space transform (8 transform classes)

5.2. Experiments on Mandarin In-Car Recognition Task

The medium vocabulary Mandarin In-car navigation command recognition system was developed using 25 hours of clean training data. A multi-style training data set was constructed by artificially corrupting the clean speech data with added car engine noise. Noise corrupted speech data generated under six sentence level SNR conditions: 0dB, 4dB, 8dB, 12dB, 16dB and 20dB, were used in training, while a corrupted 5 hour test set consists of five sentence level SNR conditions: 2dB, 6dB, 10dB, 14dB, and 18dB, was used for character error rate (CER) evaluation. The baseline acoustic models were ML trained using HTK [22] on 42-dimensional HLDA projected PLP features further augmented with smoothed pitch parameters. Phonetic decision tree clustered cross-word tonal triphones HMMs were used. A total of 2.4k distinct tied states with 12 components per state were used. A 5k word list and an associated n-gram model was used in decoding.

System	2dB	6dB	10dB	14dB	18dB	Ave
mcond baseline	44.15	27.56	20.08	17.76	17.51	25.41
tran2	51.68	31.94	23.77	18.99	17.24	28.72
tran256	40.89	27.29	21.45	17.71	16.35	24.74
mv-tran2	31.05	21.87	17.88	17.31	16.62	20.95
ftran2	51.72	33.09	23.40	22.20	17.34	29.55
ftran256	41.07	27.43	21.48	17.74	16.61	24.87
mv-ftran2	33.18	24.53	17.92	17.71	16.84	22.04
ftran256-tran256	30.75	21.42	17.70	17.54	16.11	20.70
mv-ftran2-tran2	30.29	21.80	17.91	17.26	16.32	20.72

Table 3: WER performance of baseline and GVP-HMM systems on a Mandarin in car command recognition task.

A set of experiments similar to those for Aurora 2 presented in table 2 were conducted on the In-car data. Performance of the multi-style baseline and various GVP-HMM systems, are shown in table 3. Consistent with the trend previously found in table 2, feature and model space GVP-HMM systems gave comparable error rates. The model and feature transform based GVP-HMM system, "ftran256-tran256", gave an average CER reduction of 4.71% absolute (18.5% relative) over the baseline multi-style trained "mcond" system, and the lowest average CER among all systems in the table. As shown previously in table 1, this compact "ftran256-tran256" system, (shown in the 10th line of table 1 and 8th line of table 3) used 62% fewer polynomial coefficients than the most complex GVP-HMM system, "mv-ftran2-tran2", (shown in the last line of tables 1 and 3) which simultaneously models Gaussian component, mean and feature space transform parameter trajectories. These results confirm that transform based GVP-HMMs can provide a more compact form of trajectory modelling than conventional mean and variance based VP-HMMs.

6. Conclusion

Feature space generalized variable parameter HMMs (GVP-HMM) is investigated in this paper. In addition to Gaussian means and variances and model space linear transforms, it can also compactly model the optimal trajectories of tied feature space transforms that can vary with respect to ambient noise level, with improved computational efficiency during recognition time compared with conventional MR-HMMs, VP-HMMs and mean transform based GVP-HMMs. Experimental results on Aurora 2 and a medium vocabulary Mandarin speech recognition task suggest the proposed method may be useful for noise robust speech recognition. Future research will focus on discriminative training and noise adaptive training of GVP-HMMs, and modelling multiple sources of acoustic variability.

7. References

[1] A. Bjorck & V. Pereyra (1970). "Solution of Vandermonde Systems of Equations", Mathematics of Computation (American Mathematical Society) 24(112): pp. 893-903.

[2] N. Cheng, X. Liu & L. Wang (2011). "Generalized Variable Parameter HMMs for Noise Robust Speech Recognition", in *Proc. ISCA Interspeech2011*, pp. 482-484, Florence, Italy.

[3] N. Cheng, X. Liu & L. Wang (2011). "A flexible framework for HMM based noise robust speech recognition using generalized parametric space polynomial regression" *Science China, Information Sciences*, 54(2), pp. 2481-2491, 2011.

[4] X. Cui & Y. Gong (2007). "A study of variable-parameter Gaussian mixture hidden Markov modeling for noisy speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1366-1376, 2007.

[5] A. P. Dempster, N. M. Laird & D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 39(1):1-39,1977.

[6] L. Deng, J. Droppo & A. Acero (2005). "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion." *IEEE Transactions on Speech and Audio Processing*, 13.3 (2005): pp.412-421.

[7] J. Droppo, A. Acero & L. Deng (2002). "Uncertainty decoding with SPLICE for noise robust speech recognition", in *Proc. IEEE ICASSP2002* pp. 57-60, Orlando.

[8] F. Flego & M. J. F. Gales (2009). Discriminative Adaptive Training with VTS and JUD, in *Proc. IEEE ASRU2009*, pp.170-175, Merano, Italy.

[9] K. Fujinaga, M. Nakai, H. Shimodaira & S. Sagayama (2001). "Multiple-Regression Hidden Markov Model", in *Proc. IEEE ICASSP2001*, Vol 1, pp. 513-516, Salt Lake City.

[10] M. J. F. Gales (1998). "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, 12(2):75-98, 1998.

[11] O. Kalinli, M. Seltzer, J. Droppo & Alex Acero (2010). Noise Adaptive Training for Robust Automatic Speech Recognition, *IEEE Trans. on Audio, Speech and Language Processing*, 01/2010; 18:1889-1901.

[12] D. K. Kim & M. J. F. Gales (2011). Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition.

IEEE Transactions on Audio, Speech, and Language Processing, 19(2), pp. 315-325.

[13] C.J. Leggetter & P.C. Woodland (1995). "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", *Computer Speech and Language*, 9:171-186,1995.

[14] H. Liao & M. J. F. Gales (2008). "Issues with uncertainty decoding for noise robust speech recognition", *Speech Communication*, 50:265-277, 2008.

[15] S. Lin, B. Chen & Y-M Yeh (2009). Exploring the Use of Speech Features and Their Corresponding Distribution Characteristics for Robust Speech Recognition, *IEEE Transactions on Audio Speech and Language Processing*, 17(1), pp.84-94, Jan. 2009.

[16] R. Lippmann, E. Martin & D. Paul (1987). "Multi-style training for robust isolated-word speech recognition", in *Proc. IEEE ICASSP1987*, pp. 705-708, Dallas, Texas.

[17] T. T. Kristjansson & B. J. Frey (2002). "Accounting for uncertainty in observations: A new paradigm for robust speech recognition", in *Proc. IEEE ICASSP2002*, pp. 61-64, Orlando.

[18] R. Martin (1993). "An efficient algorithm to estimate the instantaneous SNR speech signals", in *Proc. Eurospeech1993*, pp. 1093-1096, Berlin.

[19] M. Seltzer, D. Yu & Y. Wang (2013). An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition, to appear in *Proc. IEEE ICASSP2013*, Vancouver.

[20] N. B. Yoma & M. Villar (2002). Speaker Verification in noise using a stochastic version of the weighted Viterbi algorithm, *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No 3, March 2002.

[21] N. B. Yoma, I. Brito & J. Silva (2003). Language model accuracy and uncertainty in noise cancelling in the stochastic weighted Viterbi algorithm, in *Proc. Eurospeech2003*, Geneve, Switzerland, 1-4 September, 2003.

[22] S. Young et al., "The HTK Book Version 3.4.1 ", 2009.

[23] D. Yu, L. Deng, Y. Gong & A. Acero (2008). "Discriminative training of variable-parameter HMMs for noise robust speech recognition," in *Proc. ISCA Interspeech2008*, pp. 285-288, Brisbane.

[24] D. Yu, L. Deng, Y. Gong & A. Acero (2008), "Parameter clustering and sharing in variable-parameter HMMs for noise robust speech recognition," in *Proc. ISCA Interspeech2008*, pp. 1253-1256, Brisbane.

[25] D. Yu, L. Deng, Y. Gong & A. Acero (2009), "A Novel Framework and Training Algorithm for Variable-Parameter Hidden Markov Models", *IEEE Transactions on Audio, Speech and Language Processing*, Vol 17(7), pp. 1348-1360, 2009.