Generalized Variable Parameter HMMs Based Acoustic-to-articulatory Inversion

Xurong Xie[1,3]*, Xunying Liu*[1,2]*, Lan Wang*[1,3] *& Rongfeng Su*[1,3]

[1]Key Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2]Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
[3]The Chinese University of Hong Kong, Hong Kong, China

xr.xie@siat.ac.cn, xl207@cam.ac.uk, lan.wang@siat.ac.cn, rf.su@siat.ac.cn

Abstract

Acoustic-to-articulatory inversion is useful for a range of related research areas including language learning, speech production, speech coding, speech recognition and speech synthesis. HMM-based generative modelling methods and DNN-based approaches have become dominant approaches in recent years. In this paper, a novel acoustic-to-articulatory inversion technique based on generalized variable parameter HMMs (GVP-HMMs) is proposed. It leverages the strengths of both generative and neural network based modelling frameworks. On a Mandarin speech inversion task, a tandem GVP-HMM system using DNN bottleneck features as auxiliary inputs significantly outperformed the baseline HMM, multiple regression HMM (MR-HMM), DNN and deep mixture density network (MDN) systems by 0.20mm, 0.16mm, 0.12mm and 0.10mm respectively in terms of electromagnetic articulography (EMA) root mean square error (RMSE).

Index Terms: acoustic-to-articulatory inversion, generalized variable parameter HMM, deep neural network, bottleneck features

1. Introduction

Movements of articulators provide an alternative to acoustic representation of human speech. Articulatory movements generation is an important audio-visual technology. Acoustic-to-articulatory inversion predicts the articulatory movements for given speech. The precise articulatory movements used in model training and evaluation are normally recorded via electromagnetic articulography (EMA) [1]. The underlying inversion methods have been significantly improved amid the rapid progress of speech recognition and synthesis techniques in recent years. Current acoustic-to-articulatory inversion methods can be categorized into two major types.

In generative models based approaches, acoustic-to-articulatory inversion was explored using Gaussian mixture models (GMMs) in [2]. HMM-based methods jointly modelling the acoustic and articulatory data streams have also been widely used. Earlier works along this line modelled the acoustic and articulatory data streams independently [3, 4, 5]. Improved modelling of the correlation between the two via piece wise linear transformations was obtained using multiple regression HMMs (MR-HMMs) [6, 4] based models developed for articulatory speech synthesis [7]. The other category of techniques based on neural networks (NNs) directly generates the articulatory movements from the acoustic features. These include mixture density network (MDN) [8, 9, 10], and more recently deep neural networks (DNN) based techniques. In recent years, these methods have achieved state-of-the-art inversion performance [10].

An important task in acoustic-to-articulatory inversion, and machine learning in general, is to learn the optimal model structure and complexity [11, 12, 13, 14] for the underlying statistical models. For the speech inversion task, these represent the complex effects from the hidden influencing factors encoded in articulatory features on the surface acoustic realization. For generative models, a range of complexity control techniques [11, 15] have been successfully applied in related fields such as speech recognition [12, 13, 14, 16, 17, 18]. In contrast, for non-generative models such as neural networks, model selection in general remains a non-trivial problem to date.

In order to address the above issue, a novel acoustic-to-articulatory inversion technique based on generalized variable parameter HMMs (GVP-HMMs) [19, 20, 21, 22, 17, 23, 18] is proposed in this paper. It leverages the strengths of both generative and neural network based modelling frameworks. Bottleneck features [24] derived from an acoustic-to-articulatory inversion DNN [10] are used as influence factors to directly introduce controllability to the underlying GVP-HMM based generative models. The continuous EMA features space HMM parameter trajectory against these bottleneck features are modelled using polynomial functions. The optimal model structure is automatically learnt by locally optimized polynomial parameters and degrees, thus providing additional flexibility and stronger generalization than MR-HMMs. On a Mandarin speech inversion task, the proposed GVP-HMM based inversion approach significantly outperformed the baseline HMM, MR-HMM, DNN and MDN systems by 0.20mm, 0.16mm, 0.12mm and 0.10mm respectively measured in terms of EMA root mean square error (RMSE).

The rest of this paper is organized as follows. Sections 2 and 3 reviews generative models and neural network based acoustic-to-articulatory inversion methods. GVP-HMM based inversion using DNN bottleneck features is presented in sections 4 and 5. Experiments on EMA feature generation for Mandarin speech are presented in section 6. Section 7 draws the conclusions and discusses future work.

This work is supported by National Natural Science Foundation of China (NSFC 61135003, 91420301, 61401452), Shenzhen Fundamental Research Program JCYJ20130401170306806.

2. Generative model based inversion

Acoustic-to-articulatory inversion generates articulatory movements from a given acoustic representation of speech. A significant part of previous research has been focused on using generative statistical models based inversion. Among the early works, a GMM based approach models the joint distribution of acoustic and articulatory features [2]. Two-stream HMMs or their improved variants coupling the acoustic and articulatory features, have also been widely used for inversion. These two streams can be treated as independent in conventional HMMs [3, 5], or linearly correlated in MR-HMMs [6, 4], where the acoustic stream is conditioned on the articulatory features. One single Gaussian state distribution is normally used [4].

In conventional HMMs where the two streams are treated as independent, the state observation probability density function (PDF) b_q for state q at time t can be written as

$$\begin{aligned} b_q(\boldsymbol{o}_t, \boldsymbol{a}_t) &= b_q(\boldsymbol{o}_t) b_q(\boldsymbol{a}_t) \\ &= \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_q^{(o)}, \boldsymbol{\Sigma}_q^{(o)}) \mathcal{N}(\boldsymbol{a}_t; \boldsymbol{\mu}_q^{(a)}, \boldsymbol{\Sigma}_q^{(a)}) \end{aligned} \quad (1)$$

where $\mathcal{N}$ denotes a Gaussian, $\boldsymbol{o}_t$ and $\boldsymbol{a}_t$ are the acoustic and articulatory observations respectively.

In MR-HMMs [7] based inversion models, the acoustic stream is assumed to depend on the articulatory stream in the form of linear transforms, or equivalently 1st order polynomials functions. These models were originally designed to serve as generative models for speech synthesis using articulatory features. It can also be used for acoustic-to-articulatory inversion [4]. The state observation PDF is computed as

$$\begin{aligned} &b_q(\boldsymbol{o}_t, \boldsymbol{a}_t) = b_q(\boldsymbol{o}_t|\boldsymbol{a}_t) b_q(\boldsymbol{a}_t) = \\ &\mathcal{N}\left(\boldsymbol{o}_t; (\boldsymbol{\mu}_q^{(o)} + \boldsymbol{A}_q \boldsymbol{a}_t), \boldsymbol{\Sigma}_q^{(o)}\right) \mathcal{N}(\boldsymbol{a}_t; \boldsymbol{\mu}_q^{(a)}, \boldsymbol{\Sigma}_q^{(a)}) \end{aligned} \quad (2)$$

where $\boldsymbol{A}_q$ is the transform matrix representing the dependency for state q. MR-HMMs have been shown to outperform conventional HMMs based inversion. The simple linear correlation modelled between the acoustic and articulatory features is unable to fully capture the complex relationship between the two.

During training, both the acoustic and articulatory observation features including their differentials up to the 2nd order [25] are used to construct context-dependent HMM or MR-HMM phone models. In the articulatory movement generation stage, the maximum likelihood parameter generation (MLPG) algorithm [26] is used to produce static articulatory features.

3. Neural network based inversion

Current neural network based approaches often use deep neural networks (DNNs) [27, 28] based architectures. In [8] conventional DNNs used a sigmoid activation function at the output layer. The static articulatory features were used as supervised labels for training. In [8, 9, 10], deep MDN, which uses an additional GMM layer on the top of a DNN, was also proposed as a state-of-the-art DNN-based inversion technique. For the n^{th} data point, deep MDN divides the inputs to the output layer into three parts: $\boldsymbol{y}^{(\alpha)}$, $\boldsymbol{y}^{(\mu)}$ and $\boldsymbol{y}^{(\sigma)}$. These correspond to the Gaussian component weight, mean and standard deviation trajectories respectively. By using the static plus deltas and delta-deltas articulatory features $\boldsymbol{t}_n$ as supervised labels, the deep MDN is trained to minimize $E = -\sum_n \log \sum_j \mathcal{S}_j(\boldsymbol{y}^{(\alpha)}) \mathcal{N}\left(\boldsymbol{t}_n; \boldsymbol{y}_j^{(\mu)}, \exp^2(\boldsymbol{y}_j^{(\sigma)})\right)$, where $\mathcal{S}(\cdot)$ denotes the softmax function, j is the Gaussian component index, and $\mathcal{N}(\cdot)$ is a Gaussian distribution with a uniform variance for each dimension.

In common with the HMM-based methods, single Gaussian component is usually used in the GMM layer [10]. During inversion, the MLPG algorithm can be applied to generate static articulatory features as in section 2 after computing Gaussian component parameters for every frame.

4. GVP-HMM based inversion

As an alternative form of generative model based inversion method, GVP-HMMs [23] use the acoustic observation $\boldsymbol{o}_t$ as auxiliary features to generate articulatory movements directly. It assumes articulatory stream depends on acoustic stream. Equation (1) is thus re-written as

$$\begin{aligned} &b_q(\boldsymbol{a}_t, \boldsymbol{o}_t) = b_q(\boldsymbol{a}_t|\boldsymbol{o}_t) b_q(\boldsymbol{o}_t) = \\ &\mathcal{N}\left(\boldsymbol{a}_t; \boldsymbol{\mu}_q^{(a)}(\mathbf{v}_t), \boldsymbol{\Sigma}_q^{(a)}(\mathbf{v}_t)\right) \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_q^{(o)}, \boldsymbol{\Sigma}_q^{(o)}) \end{aligned} \quad (3)$$

where the trajectory functions of Gaussian means and variances of articulatory observation $\boldsymbol{a}_t$ can be represented by P order polynomials of some time auxiliary features. Therefore, it provides more flexibility than MR-HMMs in modelling the complex relationship between the articulatory and acoustic data streams. $\mathbf{v}_t^\top$ is a $(P \times N + 1)$ dimensional Vandermonde vector [29],

$$\mathbf{v}_t^\top = \left[1,\ \tilde{\mathbf{f}}_{t,1},\ ...,\tilde{\mathbf{f}}_{t,p},\ ...,\tilde{\mathbf{f}}_{t,P}\right]^\top. \quad (4)$$

and its N dimensional pth order subvector is defined as $\tilde{\mathbf{f}}_{t,p} = [v_{t,1}^p,\ ...,v_{t,j}^p,\ ...,v_{t,N}^p]^\top$, where $v_{t,j}$ is the jth element of an N dimensional factor vector Gaussian parameters are conditioned on at frame t, for example, the acoustic feature vector, $\boldsymbol{o}_t$, thus $\boldsymbol{o}_{t,j} = v_{t,j}$, or a bottleneck feature vector derived from an acoustic-to-articulatory inversion DNN,

$$\vec{\mathbf{f}}_t^{\mathsf{BN}} = [v_{t,1},\ ...,\ v_{t,j},...,\ v_{t,N}]^\top. \quad (5)$$

$\boldsymbol{\mu}_q^{(a)}(\cdot)$ and $\boldsymbol{\Sigma}_q^{(a)}(\cdot)$ are the articulatory domain P^{th} order mean and covariance trajectory polynomials of component q respectively. When diagonal covariances are used, the trajectories of the i^{th} dimension of the mean and variance parameters are

$$\begin{aligned} \mu_{q,i}^{(a)}(\mathbf{v}_t) &= \mathbf{v}_t \cdot \mathbf{c}^{(\mu_{q,i}^{(a)})} \\ \sigma_{q,i,i}^{(a)}(\mathbf{v}_t) &= \breve{\sigma}_{q,i,i}^{(a)} \mathbf{v}_t \cdot \mathbf{c}^{(\sigma_{q,i,i}^{(a)})} \end{aligned} \quad (6)$$

where $\mathbf{c}^{(\cdot)}$ is a $(P \times N + 1)$ dimensional polynomial coefficient vector and $\breve{\sigma}_{q,i,i}^{(a)}$ is the conventional HMM variance estimate.

It can be shown that the ML update solutions of the coefficient vectors can be derived as [19, 20, 23],

$$\begin{aligned} \hat{\mathbf{c}}^{(\mu_{q,i}^{(a)})} &= \mathbf{U}^{(\mu_{q,i}^{(a)})-1} \mathbf{k}^{(\mu_{q,i}^{(a)})} \\ \hat{\mathbf{c}}^{(\sigma_{q,i,i}^{(a)})} &= \mathbf{U}^{(\sigma_{q,i,i}^{(a)})-1} \mathbf{k}^{(\sigma_{q,i,i}^{(a)})} \end{aligned} \quad (7)$$

and the sufficient statistics are

$$\begin{aligned} \mathbf{U}^{(\mu_{q,i}^{(a)})} &= \sum_t \gamma_q(t) \sigma_{q,i,i}^{(a)-1}(\mathbf{v}_t) \mathbf{v}_t^\top \mathbf{v}_t \\ \mathbf{k}^{(\mu_{q,i}^{(a)})} &= \sum_t \gamma_q(t) \sigma_{q,i,i}^{(a)-1}(\mathbf{v}_t) a_{t,i} \mathbf{v}_t^\top \\ \mathbf{U}^{(\sigma_{q,i,i}^{(a)})} &= \sum_t \gamma_q(t) \breve{\sigma}_{q,i,i}^{(a)} \mathbf{v}_t^\top \mathbf{v}_t \\ \mathbf{k}^{(\sigma_{q,i,i}^{(a)})} &= \sum_t \gamma_q(t) \left(a_{t,i} - \mu_{q,i}^{(a)}(\mathbf{v}_t)\right)^2 \mathbf{v}_t^\top \end{aligned} \quad (8)$$

where $\gamma_q(t)$ is the posterior probability of frame $[\boldsymbol{a}_t, \boldsymbol{o}_t]$ being emitted from state q at time instance t.

In the above equations, the polynomial orders are assumed to be fixed at P. In order to optimize the polynomial orders, an efficient Bayesian model complexity control technique can be used [17, 23, 18]. The optimal model complexity is determined by

$$\hat{\mathcal{M}} \quad = \quad \arg\max_{\mathcal{M}} \left\{ \mathcal{Q}_{\text{ml}}^{(\mathcal{M})}(\hat{\theta}, \tilde{\theta}) - \rho \cdot \frac{k}{2} \log \mathcal{T} \right\}. \quad (9)$$

where the ML auxiliary functions $\mathcal{Q}_{\text{ml}}^{(\mathcal{M})}$ associated with Gaussian mean and variance trajectory parameters are evaluated at the optimal model parameters $\hat{\theta}$ using the statistics given in equations (7) and (8). $\mathcal{T}$ is the number of frames, k denotes the number of free parameters in model structure $\mathcal{M}$, and ρ is a tunable penalty term.

5. Articulatory inversion DNN bottleneck features for GVP-HMMs

In order to draw strengths from both generative and neural network based inversion approaches, bottleneck features [24] derived from an acoustic-to-articulatory inversion DNN [10] are used as influence factors in this paper to directly introduce controllability to the underlying GVP-HMM based generative models. This requires an acoustic-to-articulatory inversion DNN including an additional narrow bottleneck layer between the last hidden layer and the final output layer with a significantly smaller number of neurons to be constructed. This narrow layer introduces a constriction in dimensionality while retaining the useful information learned by the earlier hidden layers inside an acoustic-to-articulatory inversion DNN. An example acoustic-to-articulatory inversion DNN with a bottleneck layer used in this paper is shown in figure 1.

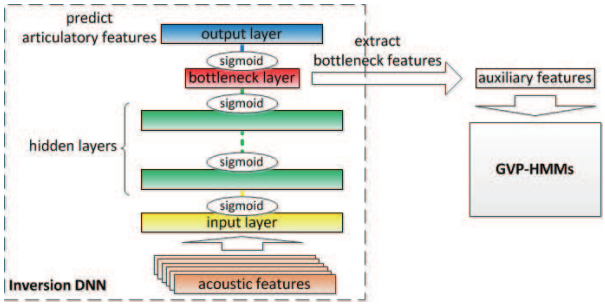

Figure 1: DNN bottleneck features for GVP-HMMs.

The resulting low dimensional DNN bottleneck features extracted from this layer are then used as auxiliary features to train GVP-HMM models of section 4. For simplicity, GVP-HMMs trained using these DNN bottleneck features are referred to as "BN GVP-HMMs" in the rest of this paper.

6. Experiments

6.1. EMA data and experiment setup

EMA [1] data can be considered to represent the articulatory movements. A data set consisting of EMA data recorded concurrently with the corresponding acoustic waveforms was used in the experiments. 3050 phonetically balanced continuous Mandarin utterances with 118 monophones were collected from a female Mandarin Chinese speaker using the same equipments and methods as the previous work in [30]. After ignoring features with small movements, a 13 dimensional static articulatory feature vector for each time instant was formed by the x-, y- and z- coordinates of seven sensors (upper and lower lip, tongue tip, tongue back, tongue dorsum, lower jaw, and right corner of the mouth (symmetric with the left)). 39 dimensional MFCC plus deltas and delta-deltas features were used as acoustic features with time shift of 0.004s, which was consistent with the sampling rate of EMA data. For the experiments, 3000 utterances with 2.8 hours of speech were used for training, and average root mean square error (RMSE) was utilized to evaluate the predicted EMA features of the remaining 50 utterances.

For the HMM-based methods, a 5-emitting-state left-to-right topology was adopted to train the two-stream triphone HMMs share-clustering to 2539 senones. The baseline two-stream independent HMMs, MR-HMMs and GVP-HMMs were trained by modified versions of HTK tools [31]. Furthermore, the state sequences for each HMM system were re-aligned one time by using the corresponding HMM system after initially generating the articulatory features. Four DNN systems were built for inversion. These include one Sigmoid DNN and one deep MDN. Both consist of 5 hidden layers, each of which had 512 neurons. In addition, two comparable bottleneck versions of these two were also built with the same configurations except for having one extra 39-node bottleneck layer before the output layer. The input of the DNNs was a context window of 11 frames 39 dimensional MFCC features selecting only every other frame. All the Sigmoid DNNs and deep MDNs were trained using a modified version of Kaldi toolkit [32]. In order to facilitate a fair comparison, all HMMs and MDNs had 1 Gaussian component, and all the GVP-HMM variance polynomial orders are fixed to 0, thus remains static.

6.2. Comparing different HMM systems for inversion

In this experiment, different two-stream HMM systems were investigated for acoustic-to-articulatory inversion. They included the MR-HMM methods with full or three-block cross-stream transform matrices [4] for each senone, and GVP-HMMs with various mean polynomial orders P and model complexity control factor ρ. The comparison of different HMM systems is presented in table 1 and figure 2, where the "Coef/State" denotes the Gaussian mean polynomial coefficients per state, and the blue bars and yellow bars indicate average RMSEs of EMA features from initial generation and from one time re-alignment generation respectively.

Sys	Model	Mean Order P	CmCtrl (ρ)	Coef/State (%)
(1)	Baseline HMM	-	-	39 (100%)
(2)	MR-HMM	1	-	1560 (100%)
(3)	MR-HMM (3-block)	1	-	546 (35%)
(4)	GVP-HMM	1	-	1560 (100%)
(5)		2	-	3081 (100%)
(6)		2	0.3	1017 (33%)
(7)		2	0.5	661 (21%)

Table 1: Configuration of different HMM systems trained using acoustic and articulatory features only.

By comparing the MR-HMMs (Sys (2)) and GVP-HMMs (Sys (4)) with mean order $P = 1$, it is clear that the average RMSE of EMA features predicted by GVP-HMM system was significantly lower than the MR-HMM systems. It indicates that the direct acoustic-to-articulatory inversion is more appropriate than inversion based on speech synthesis.

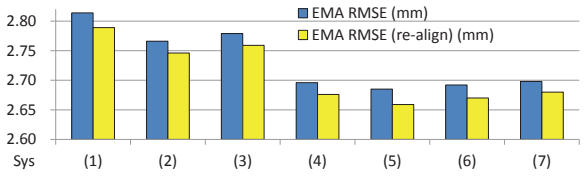

Figure 2: Performance comparison of systems in table 1 measured in EMA feature RMSE before and after re-alignment.

The GVP-HMMs (Sys (5)) with mean polynomial order $P = 2$ achieved the best performance, which was significantly lower than the baseline system (Sys (1)) by near 0.13mm EMA RMSE on both the initial and re-alignment generation respectively. Moreover, the GVP-HMMs (Sys (7)) with mean polynomial order $P = 2$ and complexity control factor $\rho = 0.5$ obtained significantly lower EMA RMSE than MR-HMMs (Sys (3)) with three-block matrices by about 0.08mm with similar model coefficients number.

6.3. Bottleneck features for GVP-HMM based inversion

In this set of experiments, acoustic-to-articulatory inversion DNN bottleneck features were used by GVP-HMM systems for EMA features generation. The performance of the four baseline DNN-based inversion methods are displayed in table 2. The use of an additional bottleneck layer led to small degradation on the resulting DNN's inversion performance. It is also worth noting that the best re-alignment GVP-HMMs trained using acoustic and articulatory features only, for example, the GVP-HMM system (5) shown in 1 and figure 2, gave a lower EMA RMSE than the best deep MDN by 0.03mm.

DNN	EMA RMSE (mm)	
	Without Bottleneck Layer	With Bottleneck Layer
Sigmoid DNN	2.711	2.741
Deep MDN	2.689	2.693

Table 2: Performance of baseline DNN-based inversion.

The above DNNs with a bottleneck layer were then utilized to extract bottleneck features for building BN GVP-HMM systems. According to the results of section 6.2, the BN GVP-HMM mean polynomial orders were fixed to 2. The results of different BN GVP-HMM systems can be found in table 3 and figure 3. It is clear that all the BN GVP-HMM systems consistently and significantly outperformed the baseline HMM, MR-HMM, Sigmoid DNN, deep MDN systems, as well as those GVP-HMM systems previously shown in table 1 trained on standard acoustic features. Furthermore, although the Sigmoid DNN performed worse than the deep MDN according to table 2, the resulting DNN BN GVP-HMM systems using its derived bottleneck features achieved a lower EMA RMSE than the comparable MDN BN GVP-HMM systems.

In addition, the re-alignment of the DNN BN GVP-HMM system (Sys (4)) achieved an EMA RMSE score of 2.590mm. This is significantly lower than the best re-alignment baseline HMM, MR-HMM, Sigmoid DNN, deep MDN systems by near 0.20mm, 0.16mm, 0.12mm, 0.10mm respectively. Additionally, the complexity control of BN GVP-HMM system significantly reduced the number of coefficients by up to 79% (line 3 in table 3), thus producing more compact model structures for BN GVP-HMMs.

A complete comparison of the above best baseline HMM, MR-HMM, Sigmoid DNN, deep MDN, GVP-HMM and DNN BN GVP-HMM systems are summarized in table 4. An example of EMA trajectory on x-axis of upper lip predicted by the systems in table 4 is also shown in figure 4.

Sys	Model	DNN	CmCtrl (ρ)	Coef/State (%)
(1)	GVP-HMM	-	-	3081 (100%)
(2)			0.3	1017 (33%)
(3)			0.5	661 (21%)
(4)	DNN BN GVP-HMM	Sigmoid DNN	-	3081 (100%)
(5)			0.3	1131 (37%)
(6)			0.5	764 (25%)
(7)	MDN BN GVP-HMM	Deep MDN	-	3081 (100%)
(8)			0.3	1099 (36%)
(9)			0.5	735 (24%)

Table 3: Different bottleneck features for GVP-HMM systems.

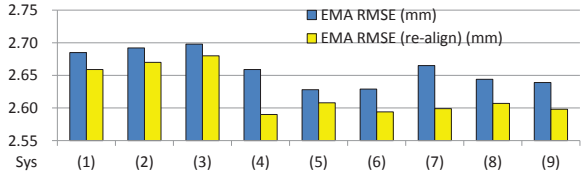

Figure 3: Comparison of systems in table 3.

System	EMA RMSE (mm)	
	Initial Generation	+Re-alignment
Baseline HMM	2.814	2.789
MR-HMM	2.766	2.746
Sigmoid DNN	2.711	-
Deep MDN	2.689	-
GVP-HMM	**2.685**	**2.659**
DNN BN GVP-HMM	**2.659**	**2.590**

Table 4: Complete comparison of the best systems

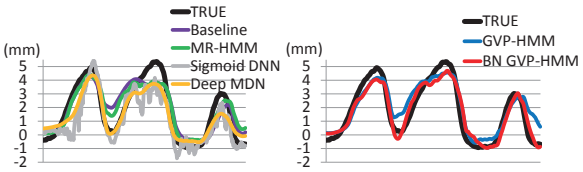

Figure 4: EMA trajectory on x-axis of upper lip. Phone sequence: *s uǒ y ǒu g òng ch ā n d ǎ ng y uá n.*

7. Conclusions

An improved acoustic-to-articulatory inversion technique based on GVP-HMMs is proposed in this paper. This method exploits the useful information from the bottleneck features derived from acoustic-to-articulatory inversion DNNs, thus can generate more precise articulatory movements. The best BN GVP-HMM system in the experiments significantly outperformed the baseline HMM, MR-HMM, Sigmoid DNN and deep MDN systems by near 0.20mm, 0.16mm, 0.12mm and 0.10mm of average EMA RMSE respectively. Future work will focus on improving the GVP-HMM based modelling architecture for speech inversion.

8. References

[1] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain Lang.*, vol. 31, pp. 26–35, 1987.

[2] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*, 2004.

[3] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[4] Z. H. Ling, K. Richmond, and J. Yamagishi, "An analysis of HMM-based prediction of articulatory movements," *Speech Communication*, vol. 52, no. 10, pp. 834–846, 2010.

[5] L. Wang, H. Chen, S. Li, and H. M. Meng, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication*, vol. 54, no. 7, pp. 845–856, 2012.

[6] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.

[7] Z. H. Ling, K. Richmond, J. Yamagishi, and R. H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[8] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, no. 2, pp. 153–172, 2003.

[9] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[10] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, 2012.

[11] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[12] X. Liu and M. J. F. Gales, "Automatic model complexity control using marginalized discriminative growth functions," in *Proc. IEEE ASRU*, St. Thomas, U.S. Virgin Islands, 2003, pp. 37–42.

[13] X. Liu, M. J. F. Gales, and P. C. Woodland, "Model complexity control and compression using discriminative growth functions," in *Proc. IEEE ICASSP*, vol. 1, Montreal, Quebec, Canada, 2004, pp. 797–800.

[14] X. Liu and M. J. F. Gales, "Automatic model complexity control using marginalized discriminative growth functions," *TRANSACTIONS*, vol. 15, no. 4, pp. 1414–1424, 2007.

[15] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[16] S. Watanabe, Y. Minam, A. Nakamura, and N. Ueda, "Variational bayesian estimation and clustering for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 365–381, 2004.

[17] R. Su, X. Liu, and L. Wang, "Automatic model complexity control for generalized variable parameter HMMs," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, 2013, pp. 150–155.

[18] ——, "Automatic complexity control of generalized variable parameters HMMs for noise robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 102–114, 2015.

[19] N. Cheng, X. Liu, and L. Wang, "Generalized variable parameter HMMs for noise robust speech recognition," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 482–484.

[20] ——, "A flexible framework for HMM based noise robust speech recognition using generalized parametric space polynomial regression," *Science China, Information Sciences*, vol. 54, no. 2, pp. 2481–2491, 2011.

[21] Y. Li, X. Liu, and L. Wang, "Structured modeling based on generalized variable parameter HMMs and speaker adaptation," in *8th International Symposium on Chinese Spoken Language Processing, ISCSLP 2012, Kowloon Tong, China, December 5-8, 2012*, 2012, pp. 136–140.

[22] ——, "Feature space generalized variable parameter HMMs for noise robust recognition," in *Proc. ISCA INTERSPEECH*, Lyon, France, 2013, pp. 2968–2972.

[23] X. Xie, R. Su, X. Liu, and L. Wang, "Deep neural network bottleneck features for generalized variable parameter HMMs," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014.

[24] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007*, 2007, pp. 757–760.

[25] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Articulatory control of hmm-based parametric speech synthesis driven by phonetic knowledge," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 2008, pp. 573–576.

[26] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2000, 5-9 June, 2000, Hilton Hotel and Convention Center, Istanbul, Turkey*, 2000, pp. 1315–1318.

[27] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30C–42, jan 2012.

[28] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kinsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, pp. 2–17, nov 2012.

[29] A. Bjorck and V. Pereyra, "Solution of vandermonde systems of equations," *Mathematics of Computation (American Mathematical Society)*, vol. 24, no. 112, pp. 893–903, 1970.

[30] D. Zhang, X. Liu, N. Yan, L. Wang, Y. Zhu, , and H. Chen, "A multi-channel/multi-speaker articulatory database in mandarin for speech visualization," in *The 9th International Symposium on Chinese Spoken Language Processing, Singapore, September 12-14, 2014*, 2014, pp. 299–303.

[31] S. Young *et al.*, *The HTK Book Version 3.4.1*, 2009.

[32] The Kaldi speech recognition toolkit, *http://kaldi.sourceforge.net*.