INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Deep Neural Network Based Acoustic-to-articulatory Inversion Using Phone Sequence Information

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

Abstract

In recent years, neural network based acoustic-to-articulatory inversion approaches have achieved the state-of-the-art performance. One major issue associated with these approaches is the lack of phone sequence information during inversion. In order to address this issue, this paper proposes an improved architecture hierarchically concatenating phone classification and articulatory inversion component DNNs to improve articulatory movement generation. On a Mandarin Chinese speech inversion task, the proposed technique consistently outperformed a range of baseline DNN and RNN inversion systems constructed using no phone sequence information, a mixture density parameter output layer, additional phone features at the input layer, or multi-task learning with additional monophone output layer target labels, measured in terms of electromagnetic articulography (EMA) root mean square error (RMSE) and correlation. Further improvements were obtained using the bottleneck features extracted from the proposed hierarchical articulatory inversion systems as auxiliary features in generalized variable parameter HMMs (GVP-HMMs) based inversion systems.

Index Terms: acoustic-to-articulatory inversion, deep neural network, bottleneck feature, phone sequence

1. Introduction

During human speech production, movements of articulators [1, 2, 3, 4] provide an important visual alternative to the acoustic representation of speech. Precise articulatory movements of both internal and external articulators are commonly recorded via electromagnetic articulography (EMA) [1]. As it is expensive to record large amounts of high quality articulatory movements, statistical inversion approaches are often used to predict articulatory movements from the acoustic data. Current articulatory inversion methods can be classified to two major types.

The first category of techniques are generative model based inversion approaches, which generally utilize hidden Markov models (HMMs) [5, 6] to jointly model the acoustic and articulatory data streams. Along this line, further improvements can be obtained by appropriately modelling the correlation between these two streams using, for example, multiple regression HMM (MR-HMM) [7, 8] and generalized variable parameter HMM (GVP-HMM) [9] based inversion approaches. One issue associated with these techniques is that during inversion the phone sequence information of utterances being processed are required. In general this can be non-trivial to obtain when the ground truth reference transcripts are not available, and the use of automatic speech recognition systems can introduce errors and significantly degrade inversion performance.

The second category is broadly based non-linear regression techniques [3, 10, 11, 12] represented by techniques based on artificial neural networks (ANNs). Early research work focuses on using conventional multi-layer perceptrons (MLPs) with a shallow hidden layer architecture [3, 10, 11]. These models were initially used to either directly map the acoustic features to articulatory movements, and later on to generate articulatory trajectory density distribution parameters using mixture density networks (MDNs) [13, 14, 15] and deep recurrent neural networks (RNNs) [16]. These techniques exploit the inherently strong generalization performance and sequence modeling power of neural networks. No phone or viseme information is used in the training stage. Instead, they are implicitly learned via the hidden layer presentations over time in an unsupervised fashion. Hence, the discrimination between adjacent acoustic or articulatory frames belonging to different phoneme or viseme units cannot be fully learned. This was found to produce noisy articulatory movement trajectories during inversion [9].

In order to address the above issue, stacked [17], hierarchical DNN and RNN based inversion approaches are proposed in this paper. A bottom level phone classification DNN or RNN taking acoustic feature inputs is used to produce bottleneck features [18]. These are in turn augmented to the conventional acoustic front-ends and fed in a top level articulatory inversion DNN or RNN network. On a Mandarin Chinese speech inversion task, the proposed technique outperformed a range of baseline deep neural network based inversion systems by statistically significant margin in terms of electromagnetic articulography (EMA) root mean square error (RMSE) and correlation. These include baseline inversion DNNs and RNNs constructed either using no phone sequence information, using additional phone features at the input layer, or multi-task learning with phone target labels also modeled at the output layer [19]. Further improvements were obtained using the bottleneck features extracted from the proposed hierarchical articulatory inversion systems as auxiliary features in generalized variable parameter HMMs (GVP-HMMs) based inversion system [9].

The rest of this paper is organized as follows. Section 2 reviews neural network based acoustic to articulatory inversion approaches. The proposed hierarchical inversion approach and their combination with GVP-HMMs are presented in sections 3

This work is supported by National Natural Science Foundation of China (NSFC 61135003, 91420301), Shenzhen Fundamental Research Program JCYJ20160601170306806.

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-659

and 4. Experiments on EMA feature generation for a Mandarin speech corpus are presented in section 5. Section 6 draws the conclusions and discusses possible future work.

2. ANN based articulatory inversion

There has been a long term research interest to exploit the inherently strong generalization and discriminative power of artificial neural networks (ANNs) for sequence modelling tasks such as articulatory inversion. Early inversion techniques explored the use of multi-layer perceptrons (MLPs) with a shallow hidden layer architecture along two related lines of research. They were used to either directly map the acoustic features to articulatory movements in the form of a conventional articulatory inversion MLPs [3, 10, 11], or used later on to model the articulatory trajectory density distribution parameters under the mixture density networks (MDNs) framework [13, 14]. With the rapid advance of deep learning techniques [20, 21] in recent years, these two related research lines have also developed into their more advanced forms. Significant inversion performance improvements have been obtained using deep neural networks (DNNs) or recurrent neural networks (RNNs) [16] based approaches, as well as the comparative deep MDNs [15, 22] based methods.

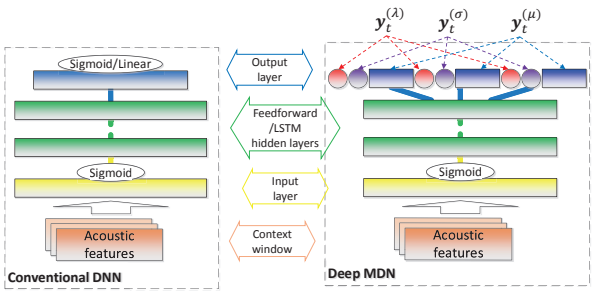

Figure 1: Examples of an articulatory inversion DNN (left) and a deep mixture density network (MDN) (right).

Examples of an articulatory inversion DNN and a deep MDN are shown in figure 1. The inputs fed into both models consist of a context window of acoustic features constructed at each time instance. The hidden layers of networks can be feedforward layers, or recurrent layers [16]. In MLP and DNN based acoustic-to-articulatory inversion neural networks, Sigmoid or linear activation functions are used at the output layer. The static articulatory features were used as supervised labels for training. In contrast, a specially designed output layer predicting GMM based articulatory trajectory parameters are used in MDNs and deep MDNs [13, 14, 15, 22]. In common with the HMM-based methods, a single Gaussian component is usually used in the GMM layer [15]. During inversion, the maximum likelihood parameter generation (MLPG) [23] algorithm can be applied to generate static articulatory features after computing Gaussian component parameters for every frame.

3. Improved ANN based articulatory inversion using phone sequence information

Several improved forms of articulatory inversion DNNs incorporating phone information are presented in this section.

3.1. Using input phone features

In this inversion architecture the standard acoustic features are augmented with binary monophone label input features before being fed into an inversion DNN or RNN as joint inputs. An example of such inversion DNN is shown in the left half of figure 2. Using this architecture, phone sequences of utterances are required in both training and testing stages.

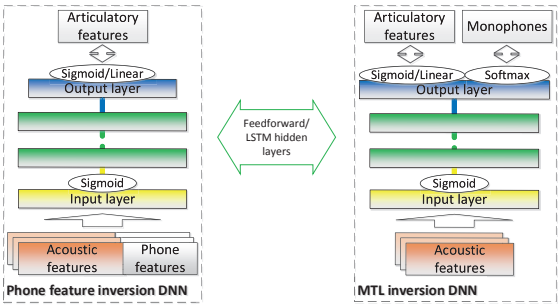

Figure 2: Examples of two articulatory inversion DNNs using phone input features (left) or multi-task learning with additional phone output targets labels (right).

3.2. Multi-task learning based inversion

Alternatively, phone sequence information can be modeled concurrently with the articulatory features using a multi-task learning (MTL) [24] trained inversion DNNs [19, 17]. In these systems, the primary task is the same as conventional articulatory inversion DNNs, while the secondary task is monophone classification using supervised labels. An example of such inversion DNN is shown in the right half of figure 2.

3.3. Hierarchical NN based articulatory inversion

The third form of improved NN inversion technique incorporating phone sequence information is based on a stacked [17], hierarchical ANN architecture. It consists of a bottom level classification sub-network and a top level inversion sub-network. The bottom level phone classification DNN or RNN sub-network taking acoustic feature as its inputs and monophone labels at its outputs is used to produce bottleneck features. The resulting bottleneck features representing the underlying phone sequence information are then augmented to the conventional acoustic features before being fed into the top level articulatory inversion DNN or RNN sub-network. An example of such hierarchical inversion DNN architecture is shown in figure 3.

The construction of a hierarchical inversion NN system involves two stages. In the first stage, the bottom level monophone classification subnetwork with a bottleneck layer is trained using the conventional cross entropy criterion. In the second stage, the bottleneck features produced by the monophone classification subnetwork are augmented to the acoustic features, and form context windows of tandem features to train the top level inversion subnetwork. Two implementation issues need to be appropriately addressed in this stage.

First, in order to ensure a fast and stable convergence during training, the top level inversion sub-network is initialized using a conventional inversion DNN or RNN trained with acoustic features as inputs only, while the newly introduced input layer weight submatrices associated with the phone classification bottleneck features are randomly initialized. Second, as only the input layer weight submatrices associated with the bottle-

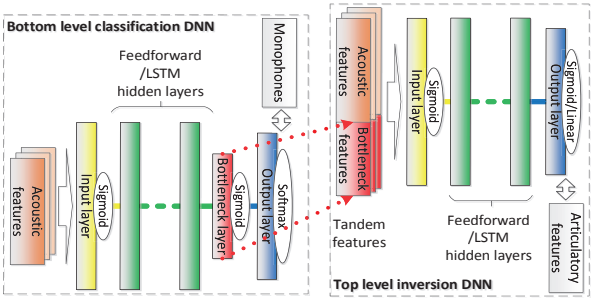

Figure 3: Example of a hierarchical inversion DNN consisting of from left to right a bottom level classification sub-network and a top level inversion sub-network

neck features require a full update while the remaining of the top level inversion sub-network only require further fine tuning, different learning rate settings are preferred for training these two parts of the top level sub-network. For implementation, in order to allow the top level inversion sub-network to more efficiently incorporate the additional information presented in the augmented bottleneck features, the actual input layer weights connecting bottleneck features are further scaled by a factor of $\alpha = 2$. This leads to an effectively larger learning rate for the the input layer weight submatrices associated with the augmented bottleneck features.

4. Hierarchical inversion bottleneck features for tandem GVP-HMM based inversion

When the phone sequences of testing utterances are known, hierarchical inversion NN systems can also be used as an auxiliary feature extractor for generalized variable parameter HMMs (GVP-HMMs) [9] based inversion systems to obtain further performance improvements. This requires an additional bottleneck layer to be added into the top level inversion sub-network before training, for example, as shown in figure 4. Given an

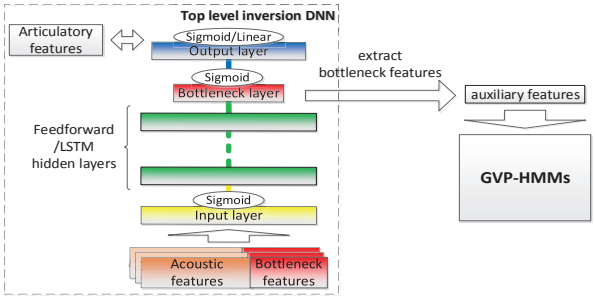

Figure 4: Example of using a hierarchical articulatory inversion DNN generated bottleneck features for training tandem GVP-HMM based inversion systems.

N dimensional bottleneck feature vector for frame t denoted by $\mathbf{f}_t^{\text{BN}} = [v_{t,1}, ..., v_{t,j}, ..., v_{t,N}]^\top$ and polynomial order P, a $(P \times N + 1)$ dimensional Vandermonde vector [25] is constructed as $\mathbf{v}_t^\top = [1, \tilde{\mathbf{f}}_{t,1}, ..., \tilde{\mathbf{f}}_{t,p}, ..., \tilde{\mathbf{f}}_{t,P}]^\top$, where its N dimensional pth order subvector is defined as $\tilde{\mathbf{f}}_{t,p} = [v_{t,1}^p, ..., v_{t,j}^p, ..., v_{t,N}^p]^\top$, and $v_{t,j}$ is the jth dimension of a bottleneck feature vector $\mathbf{f}_t^{\text{BN}}$.

This Vandermonde vector is then used to train tandem GVP-HMMs [26] that model the joint state probability density functions over both the articulatory and acoustic streams. For a given state q at tth frame, this is given by

$$b_q(\boldsymbol{a}_t, \boldsymbol{o}_t) = \mathcal{N}\left(\boldsymbol{a}_t; \boldsymbol{\mu}_q^{(a)}(\mathbf{v}_t), \boldsymbol{\Sigma}_q^{(a)}(\mathbf{v}_t)\right) \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_q^{(o)}, \boldsymbol{\Sigma}_q^{(o)})$$

where $\boldsymbol{a}_t$ and $\boldsymbol{o}_t$ denote the articulatory and acoustic features respectively.

GVP-HMMs provide a flexible modelling of the complex relationship between articulatory and acoustic data streams. The trajectory functions of Gaussian means $\boldsymbol{\mu}_q^{(a)}(\cdot)$ and variances $\boldsymbol{\Sigma}_q^{(a)}(\cdot)$ of articulatory observation $\boldsymbol{a}_t$ can be represented by P order polynomials of the given auxiliary features. When diagonal covariances are used, the trajectories of the i^{th} dimension of the mean and variance are

$$\mu_{q,i}^{(a)}(\mathbf{v}_t) = \mathbf{v}_t \cdot \mathbf{c}^{(\mu_{q,i}^{(a)})}$$
$$\sigma_{q,i,i}^{(a)}(\mathbf{v}_t) = \breve{\sigma}_{q,i,i}^{(a)} \mathbf{v}_t \cdot \mathbf{c}^{(\sigma_{q,i,i}^{(a)})},$$

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. The coefficient vectors can be estimated by maximum likelihood update scheme [27, 28, 26]. The underlying GVP-HMM model structure represented by the polynomial orders P for different parameters can be optimized using an efficient Bayesian model complexity control technique [29, 26, 30]. Given the GVP-HMM state parameter trajectories, the final articulatory movements can be generated using the MLPG algorithm.

5. Experiments

5.1. EMA data and experiment setup

Mandarin Chinese speech and associated EMA data [9, 31] were concurrently recorded by a Carstens AG-501 EMA device. After ignoring features with small movements, a 13 dimensional static EMA feature vector for each frame was chosen in order by the x- and y- coordinates of upper lip, lower lip, tongue back, tongue dorsum, tongue tip and lower jaw, and the z- coordinates of right (symmetric with the left) corner of the mouth, and further normalized by setting the average static position to zero. 39 dimensional MFCC acoustic features were extracted from the speech waveforms. The average root mean square error (RMSE) and correlation metrics were utilized to evaluate the predicted EMA data. The training, validation and evaluation sets contain 2950, 50 and 50 utterances respectively, with about 2.9 hours in total.

When constructing various hierarchical inversion ANNs, component DNN or deep RNN sub-networks were trained for monophone classification and articulatory inversion respectively. The articulatory inversion DNNs consisted of 5 feedforward hidden layers with 512 neurons. The comparable inversion RNNs consisted of 3 feed-forward hidden layers with 512 neurons each and 2 LSTM layers with 128 cells and recurrent projection layers for dimensionality reduction [32]. A context window of 11-frame MFCCs selecting only every other frame was used as their input, and the 13 static EMA features were used as the targets. The classification DNN and RNNs with a bottleneck layer were trained using the same input acoustic features while taking the 118 tonal monophone labels as the output targets. These were then used to produce 39 dimensional bottleneck features that were concatenated with the acoustic

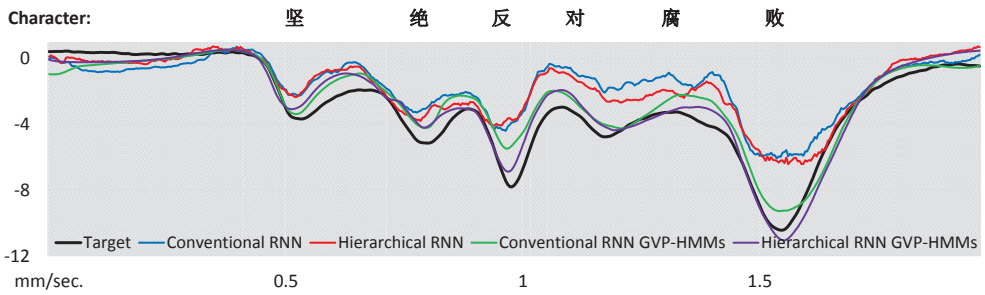

Figure 5: An example of EMA trajectory on y-axis of lower jaw predicted by inversion systems in table 1 and 2 for a Mandarin Chinese speech segment with the phoneme sequence: *j iā n j ué f ǎ n d uì f ǔ b ài*.

features to train the top level inversion DNN or RNN subnetworks. Phone input feature based and multi-task learning inversion ANNs described in sections 3.1 and 3.2 as well as MDN based inversion systems were also trained for comparison. Finally, a bottleneck version of the best performing hierarchical inversion system was used to produce auxiliary features for a tandem GVP-HMM inversion system to achieve further performance improvements. A modified version of the Kaldi toolkit [33] and HTK tools [34] were used to train all the neural networks and GVP-HMMs respectively in the experiments. For all results presented in this paper, paired t-test based statistical significance tests were performed on the RMSE reduction.

5.2. Experiments results

The inversion performance of various baseline and hierarchical DNN or RNN based inversion systems measured in term of the predicted EMA features' RMSE and correlation scores are shown in table 1, where p-value column shows only the RMSE reduction significance compared with the corresponding "Phn" NN baseline. A general trend can be found for both the DNN and RNN based inversion systems. Their respective hierarchical inversion systems (line 8 and 14 in table 1) statistically significantly outperformed the baseline inversion DNNs (line 1 to 4 in table 1), and RNNs (line 9 to 12 in table 1), constructed either using no phone sequence information, a Gaussian mixture density parameter based output layer, additional phone features at the input layer, or multi-task learning with phone target labels also modeled at the output layer. As expected, the RNN based bottom level phone classification sub-network is able to retain longer temporal information from the input phone sequence than a DNN based phone classifier. This advantage is also shown in their respective inversion performance after producing bottleneck features for the top level inversion DNN (line 7, 8 in table 1). Using RNNs as both the bottom level phone classification sub-network and the top level inversion sub-network (line 14 in table 1) gave the best inversion performance among all systems in table 1. This hierarchical RNN inversion system gave an RMSE score of 2.411 and a correlation score of 0.768.

	Inversion NN	Monophone Classifi. NN	RMSE (mm)	Correlation	p-value
1.	DNN	-	2.711	0.684	-
2.	deep MDN		2.688	**0.743**	
3.	Phn DNN		2.633	0.700	
4.	MTL DNN		2.717	0.676	
5.	Hier DNN*	DNN	2.689	0.702	-
6.		RNN	2.623	0.727	0.108
7.	Hier DNN	DNN	2.669	0.692	-
8.		RNN	**2.588**	0.724	0.000
9.	RNN	-	2.493	0.752	-
10.	recur. MDN		2.460	0.765	
11.	Phn RNN		2.449	0.759	
12.	MTL RNN		2.492	0.751	
13.	Hier RNN*	RNN	2.437	0.760	0.083
14.	Hier RNN	RNN	**2.411**	**0.768**	0.000

Table 1: Inversion performance of various baseline and hierarchical DNN and RNN based inversion systems measured in term of predicted EMA features' RMSE and correlation scores (* Without conventional NN for top level sub-network initialization).

The inversion performance of tandem GVP-HMM inversion systems trained using bottleneck features produced by baseline DNN and RNN inversion systems (line 1 and 9 in table 1) and their respective comparable hierarchical systems (line 8 and 14 in table 1) constructed with an additional bottleneck layer. As expected, further small but consistent improvements in the RMSE and correlation scores were obtained using tandem GVP-HMM based inversion systems trained on the bottleneck features produced by these systems.

	Inversion BN NN	RMSE (mm)	Correlation	p-value
1.	DNN	2.594	0.753	-
2.	Hier DNN	**2.553**	**0.763**	0.000
3.	RNN	2.422	0.760	-
4.	Hier RNN	**2.378**	**0.768**	0.004

Table 2: Inversion performance of tandem GVP-HMM systems trained using bottleneck features produced by baseline and hierarchical DNN and RNN inversion systems in table 1 with an additional bottleneck layer.

Figure 5 shows an example EMA trajectory predicted by conventional and hierarchical inversion RNN systems in table 1 and their corresponding tandem GVP-HMM systems in table 2 respectively for a Mandarin Chinese speech segment.

6. Conclusion

A hierarchical neural network based articulatory inversion architecture is proposed in the paper. On a Mandarin Chinese speech inversion task, the proposed technique was found to generate consistently more precise articulatory movements than the baseline DNN or RNN based inversion systems constructed using no phone sequence information, a mixture density parameter based output layer, additional phone features at the input layer, or multi-task learning with phone target labels also modeled at the output layer. Experimental results suggests the proposed technique may be useful for articulatory inversion and articulatory speech synthesis. Future work will focus on improving its generalization and adaptation to mismatched speakers.

7. 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 and Language*, vol. 31, no. 1, pp. 26–35, 1987.

[2] T. Baer, J. Gore, S. Boyce, and P. Nye, "Application of mri to the analysis of speech production," *Magnetic Resonance Imaging*, vol. 5, no. 1, pp. 1–7, 1987.

[3] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *Journal of the Acoustical Society of America*, vol. 92, pp. 688–700, Aug. 1992.

[4] J. R. Westbury, "X-ray microbeam speech production database user's handbook (version 1.0)," Waisman Center, University of Wisconsin-Madison, pp. 1–135, 1994.

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

[6] 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.

[7] 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.

[8] Z. 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.

[9] X. Xie, X. Liu, L. Wang, and R. Su, "Generalized variable parameter hmms based acoustic-to-articulatory inversion," in *ISCA Interspeech*, Dresden, Germany, September 6-10, 2015, pp. 279–283.

[10] B. S. Atal and O. Rioul, "Neural networks for estimating articulatory positions from speech," *Journal of the Acoustical Society of America*, vol. 86, no. S1, pp. S67–S67, 1989.

[11] M. Rahim, C. Goodyear, B. Kleijn, J. Schroeter, and M. Sondhi, "On the use of neural networks in articulatory speech synthesis," *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1109–1121, 1993.

[12] H. Kjellström and O. Engwall, "Audiovisual-to-articulatory inversion," *Speech Communication*, vol. 51, no. 3, pp. 195–209, 2009.

[13] 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.

[14] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *ISCA ICSLP*, Pittsburgh, PA, USA, September 17-21, 2006.

[15] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *ISCA Interspeech*, Portland, Oregon, USA, September 9-13, 2012, pp. 867–870.

[16] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *IEEE ICASSP*, Brisbane, Australia, April 19-24, 2015, pp. 4450–4454.

[17] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *IEEE ICASSP*, Brisbane, Australia, April 19-24, 2015, pp. 4460–4464.

[18] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *IEEE ICASSP*, Honolulu, Hawaii, USA, April 15-20, 2007, pp. 757–760.

[19] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," in *IEEE ICASSP*, Brisbane, Australia, April 19-24, 2015, pp. 4290–4294.

[20] 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. 30–42, January 2012.

[21] 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, November 2012.

[22] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *IEEE ICASSP*, Florence, Italy, May 4-9, 2014, pp. 3844–3848.

[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE ICASSP*, Istanbul, Turkey, 5-9 June, 2000, pp. 1315–1318.

[24] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[25] 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.

[26] X. Xie, R. Su, X. Liu, and L. Wang, "Deep neural network bottleneck features for generalized variable parameter HMMs," in *ISCA Interspeech*, Singapore, September 14-18, 2014, pp. 2739–2743.

[27] N. Cheng, X. Liu, and L. Wang, "Generalized variable parameter HMMs for noise robust speech recognition," in *ISCA Interspeech*, Florence, Italy, August 27-31, 2011, pp. 482–484.

[28] ——, "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.

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

[30] ——, "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.

[31] 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 *IEEE ISCSLP*, Singapore, September 12-14, 2014, pp. 299–303.

[32] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *ISCA Interspeech*, Singapore, September 14-18, 2014, pp. 338–342.

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

[34] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book Version 3.4.1,* Cambridge University Engineering Department. Cambridge University Engineering Department, 2009.