Interspeech 2018
2-6 September 2018, Hyderabad

Semi-supervised Cross-domain Visual Feature Learning for Audio-Visual Broadcast Speech Transcription

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

[1]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences
[3]The Chinese University of Hong Kong, Hong Kong, China

rf.su@siat.ac.cn, xyliu@se.cuhk.edu.hk, lan.wang@siat.ac.cn

Abstract

Visual information can be incorporated into automatic speech recognition (ASR) systems to improve their robustness in adverse acoustic conditions. Conventional audio-visual speech recognition (AVSR) systems require highly specialized audio-visual (AV) data in both system training and evaluation. For many real-world speech recognition applications, only audio information is available. This presents a major challenge to a wider application of AVSR systems. In order to address this challenge, this paper proposes a semi-supervised visual feature learning approach for developing AVSR systems on a DARPA GALE Mandarin broadcast transcription task. Audio to visual feature inversion long short-term memory neural networks (L-STMs) were initially constructed using limited amounts of out of domain AV data. The acoustic features domain mismatch against the broadcast data was further reduced using multi-level domain adaptive deep networks. Visual features were then automatically generated from the broadcast speech audio and used in both AVSR system training and testing time. Experimental results suggest a CNN based AVSR system using the proposed semi-supervised cross-domain audio-to-visual feature generation technique outperformed the baseline audio only CNN ASR system by an average CER reduction of 6.8% relative. In particular, on the most difficult Phoenix TV subset, a CER reduction of 1.32% absolute (8.34% relative) was obtained.

Index Terms: audio-visual speech recognition, semi-supervised, visual feature learning, domain adaptation

1. Introduction

Visual information can be incorporated into automatic speech recognition (ASR) systems to improve their robustness in adverse acoustic conditions. The use of visual features in audio-visual speech recognition (AVSR) systems is motivated by the bimodal speech generation mechanism [1, 2] and the ability of humans to better distinguish spoken sounds when both audio and video are available [3]. Additionally, the visual features that are invariant to acoustic signal corruption can provide complementary information to the speech recognizer.

In recent years, various AVSR modeling techniques [4, 5, 6, 7, 8, 9, 10] have been developed and yielded an impressive improvement over the ASR systems using only audio in an adverse environment. Conventional AVSR systems based on these approaches require highly specialized audio-visual (AV) data in both system training and evaluation. However, for many real-world speech recognition applications, since only audio information is available, conventional AVSR modeling techniques are difficult to be applied. This presents a major challenge to a wider application of AVSR systems.

This paper aims to construct AVSR systems on audio-only data in real life, together with limited amounts of AV data, which is often obtained in a constrained environment and thus out of the domain of such audio-only data. Earlier works along this line used inversion models [11, 12, 13] to produce the simulated articulatory features from speech signals for improving the system performance. However, the acoustic feature domain mismatch in the inversion model training and inference stages is not discussed in these works. If the inversion models' nature is domain-specific, such domain mismatch will lead to the generation of unreliable visual features that can degrade AVSR system performance.

To address this issue, this paper proposes a novel semi-supervised cross-domain visual feature learning approach for developing AVSR systems on a typical real-world audio-only speech recognition task - a DARPA GALE Mandarin broadcast transcription task. In this approach, to handle multiple speaker data sets and adverse acoustic conditions in practical applications, a small multi-speaker 3D AV data set [14] containing far-field recordings is used for AV inversion model training; audio-to-visual inversion long short-term memory neural network (L-STM) models were initially trained using such limited out of domain AV data; the acoustic features domain mismatch against the broadcast data was further reduced using multi-level domain adaptive deep networks; visual features were then automatically generated from the broadcast speech audio and used in both AVSR system training and testing time. The proposed method therefore allows a wider application of AVSR techniques to many practical situations, when only in-domain audio data and out-of-domain AV data are available.

The rest of this paper is organized as follows. Section 2 describes the 3D AV data set. The audio-to-visual feature learning approaches are reviewed in Section 3. Section 4 proposes the semi-supervised cross-domain visual feature learning approach. The AVSR system architecture is presented in Section 5. Experiments and results are reported in Section 6. Section 7 draws the conclusions and future works.

2. 3D audio-visual data

The multi-speaker 3D AV data set [14] contains Mandarin Chinese audio and visual speech recordings. The audio data was recorded on both near-field (mouth-to-microphone distance of 10 cm) and far-field (mouth-to-microphone distance of 80 cm) conditions with 16 kHz sampling rate, 16 bit encoding, and single channel. The training set consists of 10 males and 10 females with 19.6 hours. The development set includes 4 males

10.21437/Interspeech.2018-1063

and 4 females with 5 hours. Both training and development sets include near-field and far-field audio data. The corresponding visual data was the 3D positions of 37 reflective markers on the speakers' faces and 4 headband markers, which was captured by a commercially system called OptiTrack[1] with six infrared cameras at 100 fps. The acquired AV data were asynchronous, since they were recorded on different machines. In each take, the recording of 3D visual data only happened between the starting and ending signals, and such signals were also recorded by the near-field and far-field microphones. Thus, the synchronization of AV data can be addressed by aligning the offset of the corresponding signal in each audio file.

Among the acquired 3D movements of 41 markers, this paper only used the information of 8 markers around the lip. The following 5 steps were used to obtain speaker-independent visual features: (1) remove the global head motions from the acquired facial motion data of 41 markers using the rotation and translation matrices [14], which were estimated based on the fixed distances among 4 headband markers during recording; (2) extract the motion data of lip points $\boldsymbol{l}_{i,t}$, where $\boldsymbol{l}_{i,t}$ is a 24-dimensional (8*3=24) vector of i-th utterance at time instance t; (3) remove the static lip using $\hat{\boldsymbol{l}}_{i,t} = \boldsymbol{l}_{i,t} - \boldsymbol{l}_{i,static}$, where the static lip $\boldsymbol{l}_{i,static}$ is simply selected from the first frame of the i-th utterance; (4) apply the speaker-level normalization of zero mean and unit variance to $\hat{\boldsymbol{l}}_{i,t}$; (5) apply principal component analysis (PCA) to those results obtained from step (4), and reduce the dimensionality from 24 to 16.

3. Audio-to-visual feature learning

There are two categories of audio-to-visual feature learning or inversion techniques: HMM based [15, 16, 17] inversion approaches and neural network based non-linear mapping approaches [18, 19, 20]. Since the AV relationship is highly nonlinear [21], neural network based approaches are preferred to be used for the AV inversion modeling. Moreover, in order to incorporate long-range temporal correlation present in AV data, recurrent neural network (RNN) based inversion models can be used. Among these, long short-term memory neural networks (LSTMs) [22], which is a special kind of RNNs, has been successfully applied in acoustic-to-articulatory inversion tasks [23, 24]. Inspired by its success, LSTM based methods are used for learning visual features from acoustic features.

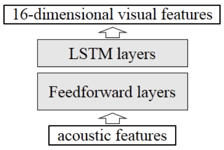

Figure 1: *Inversion LSTM: audio-to-visual feature learning.*

An example of LSTM based audio-to-visual inversion approaches used in this paper is shown in Figure 1. The inputs were a context window of 11 frames constructed at every other time instance and the targets were 16-dimensional visual features obtained in Section 2. As we know, if all layers in Figure 1 are LSTM layers, long inputs would lead to considerably longer training time and larger latency during testing. To handle such efficiency issue, we used the inversion LSTM architecture containing 3 fully connected feedforward layers with 1024 nodes each and sigmoid activation followed by 2 LSTM layers with 128 cells each. This is found in practice to give a good trade-off between performance and inference time. The fully connected feedforward layers were supposed to extract more representative acoustic features associated with visual features. Each LSTM layer contained a recurrent projection layer with 64 units for the dimensionality reduction [25], which was used to capture the dynamic long-term dependencies between acoustic and visual features.

The training criterion of inversion LSTM models is MSE based regression error. RBM based pretraining [26] was used to initialize the weights of feedforward layers. And then each LSTM layer was trained by an incremental layer-wise method [27]. When adding a new LSTM layer, the previous output weights were discarded and new random output weights were used to connect the new top layer. Note that the difference from the common layer-wise pretraining method is that only the added layer was trained at a time. Finally, all network weights were fine-tuned.

4. Semi-supervised cross-domain visual feature learning

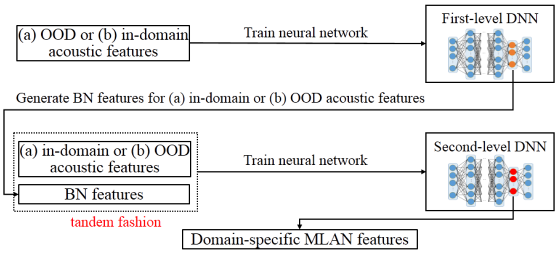
First-level DNN
(a) OOD or (b) in-domain acoustic features
Train neural network
Generate BN features for (a) in-domain or (b) OOD acoustic features
Second-level DNN
(a) in-domain or (b) OOD acoustic features
Train neural network
BN features
tandem fashion
Domain-specific MLAN features

Figure 2: *Multi-level domain adaptive deep network (MLAN): two domain adaptation directions. (a) OOD → in-domain ("3D AV" to "GALE" in this paper): generating in-domain MLAN features for AV data's audio to construct inversion model directly transforming GALE audio into visual features. (b) in-domain → OOD ("GALE" to "3D AV" in this paper): generating OOD MLAN features for in-domain GALE audio data to use OOD AV data trained inversion model.*

In contrast to the out of domain (OOD) AV data, the real-world audio-only data can be considered as in-domain data. As discussed before, considering the inversion model's nature as domain-specific, the domain mismatch between the AV data's audio data and audio-only data in real life needs to be appropriately handled by multi-level domain adaptive deep network (MLAN) before the visual feature inversion can be reliably performed. A MLAN can be seen as stacked deep neural networks (DNNs), which is firstly proposed for the cross-domain adaptation [28, 29] from OOD to in-domain. Moreover, an "in-domain to OOD" MLAN was also investigated in this paper to assess the effect of acoustic perturbation on the quality of generated visual features, assuming the cleaner the acoustic condition, more reliable the visual features. Take the MLAN in Figure 2 with case (a) as an example: the bottleneck (BN) features derived from the first DNN are supposed to contain the OOD information; using such BN features as the inputs of the second DNN enables it to cope with OOD information; moreover, the second DNN is supposed to have the ability to extract the most discriminative elements associated with in-domain from OOD information, since it is trained on the in-domain audio data. Thus, the MLAN features derived from the BN layer of the second DNN

[1] http://www.naturalpoint.com/optitrack/

are domain-specific features associated with the target domain. In this paper, the 3D AV data described in Section 2 was used as OOD data, while the DARPA GALE broadcast speech audio was used as in-domain data. Each DNN in a MLAN contained 6 fully connected feedforward layers with 2048 nodes each followed by one BN layer with 39 nodes, and each layer contained sigmoid activation function. The outputs of each DNN were 118 mono phones.

Taking the advantages of MLANs and LSTM based audio-to-visual feature inversion techniques, a semi-supervised cross-domain visual feature learning approach is proposed in this paper. The inversion cross-domain LSTM (CDLSTM) models based on such approach (see Figure 3) trained on limited OOD AV data are designed for learning reliable visual features from the real-world audio-only data in both AVSR system training and evaluation. As shown in Figure 3, the standard acoustic features are concatenated with the derived MLAN features in a tandem fashion to form the new features. The advantages of using such domain-specific tandem features as inversion LSTM inputs in Figure 3 are: on the one hand, the complex relationship between audio and visual data can be maintained in the original standard acoustic features; on the other hand, the domain mismatch in acoustic space can be addressed by the MLAN features, which are associated with the same domain. Besides, this paper will explore the effect of using different standard acoustic features (PLP/MFCC/FBANK) in the inversion models on the final AVSR system performance, which will be discussed in details in the experiments.

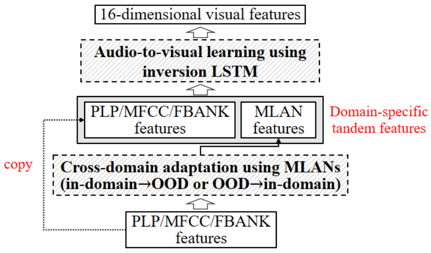
16-dimensional visual features
Audio-to-visual learning using inversion LSTM
PLP/MFCC/FBANK features
MLAN features
Domain-specific tandem features
copy
Cross-domain adaptation using MLANs (in-domain→OOD or OOD→in-domain)
PLP/MFCC/FBANK features

Figure 3: *Inversion cross-domain LSTM (CDLSTM): semi-supervised cross-domain visual feature learning.*

5. AVSR System Architecture

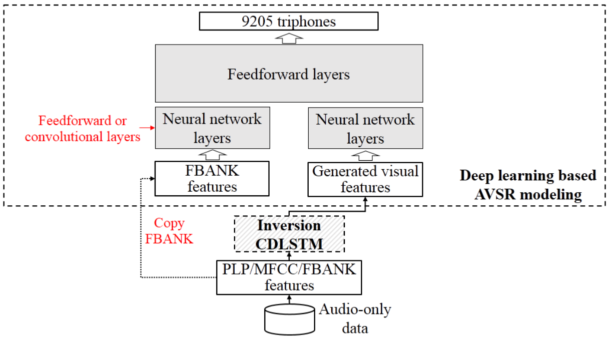

Figure 4: *AVSR system architecture using inversion CDLSTM models on real-world audio-only data (GALE broadcast speech audio data in this paper).*

Figure 4 shows the AVSR system architecture using the proposed semi-supervised cross-domain visual feature learning approach described in Section 4 on the audio-only data in real life. Since deep learning frameworks have shown impressive performance in AVSR tasks [7, 8], deep learning based audio-visual integration methods are used in this paper. As shown in Figure 4, we use separated layers in the deep learning based AVSR modeling, which are used to represent the AV inputs into the same feature space, which becomes easier for the top layers to learn higher-order correlations between audio and visual data. Especially, we explored two different types of AVSR modeling for the GALE broadcast speech transcription task.

Deep neural network (DNN) based AVSR modeling: it contained 2 separated fully connected feedforward layers for audio (2048 nodes in each hidden layer) and visual inputs (512 nodes in each hidden layer) respectively, followed by 4 fully connected feedforward layers with 2048 nodes each.

Deep convolutional neural network (CNN) based AVSR modeling: it contained 2 separated convolutional layers for audio and visual inputs respectively, followed by 4 fully connected feedforward layers with 2048 nodes each. For both audio and visual parts, the max pooling strategy was used in the first convolutional layers (256 hidden units, filter size of 6 and filter shift of 1), while no pooling techniques were used in the second convolutional layers (64 hidden units, filter size of 4 and filter shift of 1). Full weight sharing (FWS) approach [30] and 1-D feature maps along the time axis were used.

For all DNN/CNN based AVSR systems, sigmoid functions were used in each non-convolutional layer. The acoustic inputs were 11 successive frames of 72-dimensional FBANK features (static, Δ and $\Delta\Delta$) and the visual inputs were 11 successive frames of 48-dimensional visual features (static, Δ and $\Delta\Delta$). Discriminative pretraining followed by global fine-tuning using the minimum cross-entropy criterion was also performed.

One key issue associated with the proposed AVSR modeling approach using semi-supervised cross-domain visual feature learning and many other unsupervised learning techniques in general, is the reliability and appropriate selection of the generated data. In this paper, we assume all generated visual features from audio-only data will be used for subsequent AVSR training and evaluation. The data selection issue is beyond the scope of this paper and will be investigated in our future research. As shown in Figure 4, using the inversion CDLSTM models, the proposed DNN/CNN based AVSR systems use audio-only data as inputs. The only part of the AVSR systems requiring AV data is the inversion model training stage. The proposed approach allows a wider use of AVSR systems in the real world, since only audio data is required at test time.

6. Experiments

For the DARPA GALE Mandarin Chinese broadcast speech recognition task, we used GALE Phase 2 Chinese Broadcast Conversation Speech (LDC2013S04), GALE Phase 2 Chinese Broadcast News Speech (LDC2013S08), GALE Phase 3 Chinese Broadcast Conversation Speech Part 1 (LDC2014S09) and the associated transcripts (LDC2013T08, LDC2013T20, LDC2014T28), totally about 200 hours with 29 shows and 506 episodes. Among these, the development set was selected from 10% latest episodes of each show, while the rest speech were used as the training set. The test set was formed by the combination of the 07 development set and 07 evaluation set, totally 4.7 hours with 26 shows, 157 episodes, and 3170 utterances. This test set was divided into three subsets: the official TV subset, which contained standard Mandarin speech; the NTD TV subset, which was the data did not appear in the training set; the

Phoenix TV subset, which consisted of the conversational and spontaneous speech. ANN-HMM hybrid framework [31] was used in the decoding. A GMM-HMM baseline with 9205 tied states was obtained using HTK [32] and used as the bootstrap model, while all neural networks were then trained and evaluated using Kaldi toolkit [33]. The system performance was evaluated by character error rate (CER).

The supervised labels were 9205 triphone states, which were fixed by the realignment using a well-trained DNN using audio-only data. The baseline DNN ASR system contained 6 fully connected feedforward layers with 2048 nodes each. The baseline CNN ASR system contained 2 convolutional layers followed by 4 fully connected feedforward layers, and the convolutional layer configurations are the same as those of the audio part in CNN based AVSR systems described in Section 5. Besides, two more competitive DNN and CNN ASR baselines using stacked BN features [34] (2nd and 6th line in Table 2) were also obtained. The stacked DNN/CNN ASR baseline contained two DNNs/CNNs, where the first DNN/CNN containing an additional BN layer with 39-dimensional nodes before the output layer was used to extract BN features and the second DNN/CNN with such BN features in a tandem fashion as inputs was used for computing the final state posterior probabilities.

Table 1: *Average CER performance of the proposed DNN/CNN based AVSR systems using various acoustic features as inversion CDLSTM model inputs.*

Systems	Inversion CDLSTM model		CER (%)
	Features	MLAN target domain	
DNN AVSR	PLP	3DAV (fig. 2b)	11.85
		GALE (fig. 2a)	11.79
	MFCC	3DAV (fig. 2b)	11.54
		GALE (fig. 2a)	11.50
	FBANK	3DAV (fig. 2b)	11.51
		GALE (fig. 2a)	11.40
CNN AVSR	PLP	3DAV (fig. 2b)	11.56
		GALE (fig. 2a)	11.33
	MFCC	3DAV (fig. 2b)	11.30
		GALE (fig. 2a)	11.23
	FBANK	3DAV (fig. 2b)	11.16
		GALE (fig. 2a)	**11.10**

Table 1 can be partitioned into two parts. The first and second parts show the average CER performance of various proposed DNN and CNN based AVSR systems respectively. Compared with those results of using PLP and MFCC features as inversion CDLSTM model inputs, both DNN and CNN based AVSR systems using FBANK features have relative low CER performance. When the AVSR modeling types and the acoustic features used in inversion CDLSTM models were fixed in Table 1, small CER reductions were obtained from the AVSR systems using MLAN features associated with GALE domain over those associated with 3DAV domain. Besides, using the proposed semi-supervised visual feature learning techniques, CNN based AVSR systems outperformed the comparable DNN based AVSR systems. For example, an average CER reduction of 0.30% absolute (2.6% relative) was obtained from the CNN based AVSR system in the last line of the second part of Table 1 over the corresponding DNN based AVSR system in the last line of the first part of Table 1.

In order to better evaluate the proposed semi-supervised cross-domain visual feature learning approach in AVSR modeling, Table 2 shows detailed CER performance of the baseline DNN/CNN ASR systems with/without stacked BN features and DNN/CNN based AVSR systems using inversion LSTM and inversion CDLSTM with FBANK feature as inputs. Compared with those results of the ASR baseline in the first line of the first/second part of Table 2, no CER performance improvement was obtained from the DNN/CNN based AVSR systems directly using visual feature learning approaches without MLAN domain adaptation techniques. For example, an average CER increase of 0.57% absolute was obtained from the DNN AVSR system using inversion LSTM model in the 3rd line of Table 2 over the DNN ASR baseline in the first line of Table 2. As expected, use semi-supervised cross-domain visual feature learning approach in AVSR modeling gave the reductions in CER. For example, the CNN AVSR system using inversion CDLSTM with the MLAN adaptation to GALE domain (highlighted in bold in Table 2) outperformed the comparable CNN ASR baseline by an average CER reduction of 6.8% relative, and it also outperformed the more complex CNN ASR baseline using stacked BN features by an average CER reduction of 4.3% relative. Specially, on the most difficult Phoenix TV subset, a CER reduction of 1.32% absolute (8.34% relative) was obtained over the CNN ASR baseline.

Table 2: *Details of CER performance of the baseline DNN/CNN ASR systems, DNN/CNN based AVSR systems using inversion LSTM and inversion CDLSTM with FBANK feature as inputs. (The MLAN in Figure 2a was used in inversion CDLSTM).*

Systems	BN feats.	AV inversion model	CER (%)			
			Offi.	NTD.	PHNX.	Avg.
DNN	×	×	10.65	10.63	16.95	12.77
ASR	a	×	9.99	10.15	16.16	12.08
DNN	a+v	LSTM(fig. 1)	11.00	11.07	17.60	13.34
AVSR		CDLSTM(fig. 3)	9.65	9.56	15.11	11.40
CNN	×	×	9.93	9.94	15.83	11.91
ASR	a	×	9.83	10.08	15.01	11.60
CNN	a+v	LSTM(fig. 1)	10.12	10.14	16.00	12.14
AVSR		CDLSTM(fig. 3)	**9.35**	**9.49**	**14.51**	**11.10**

7. Conclusions

In this paper, a semi-supervised cross-domain visual feature learning approach was proposed for constructing audio-visual speech recognition (AVSR) systems on a DARPA GALE Mandarin broadcast transcription task. Compared with the conventional AVSR systems, the proposed AVSR systems can be used when only audio information is available in many practical applications. Experimental results suggest that, using the proposed cross-domain audio-to-visual feature generation techniques in both system training and testing, a CNN based AVSR system outperformed the baseline audio only CNN ASR system by an average CER reduction of 6.8% relative. In particular, on the most difficult Phoenix TV subset, a CER reduction of 1.32% absolute (8.34% relative) was obtained. Future work will focus on the data selection of generated visual features.

8. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC U1736203), ShenZhen Fundamental Research Program JCYJ20160429184226930 and KQJSCX20170731163308665, MSRA collaborative research grant no. 6904412, and Chinese University of Hong Kong direct allocation grant no. 4055065.

9. References

[1] B. Dodd and R. Campbell, *Hearing by eye: The psychology of lip-reading*. Lawrence Erlbaum Press, 1987.

[2] D. W. Massaro and M. M. Cohen, "Evaluation and integration of visual and auditory information in speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 9, no. 5, pp. 753–771, 1983.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[4] D. Ivanko, A. Karpov, D. Ryumin, I. Kipyatkova, A. Saveliev, V. Budkov, D. Ivanko, and M. Železnỳ, "Using a high-speed video camera for robust audio-visual speech recognition in acoustically noisy conditions," in *Proc. SPECOM*, 2017, pp. 757–766.

[5] Y. Miao and F. Metze, "Open-domain audio-visual speech recognition: A deep learning approach." in *Proc. ISCA INTERSPEECH*, 2016, pp. 3414–3418.

[6] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. IEEE ICASSP*, 2015, pp. 2130–2134.

[7] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, and K. Takeda, "Integration of deep bottleneck features for audio-visual speech recognition," in *Proc. ISCA INTERSPEECH*, 2015, pp. 689–696.

[8] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[9] J. Wang, J. Zhang, K. Honda, J. Wei, and J. Dang, "Audio-visual speech recognition integrating 3D lip information obtained from the kinect," *Multimedia Systems*, vol. 22, no. 3, pp. 315–323, 2016.

[10] A. Thanda and S. M. Venkatesan, "Multi-task learning of deep neural networks for audio visual automatic speech recognition," *arXiv:1701.02477*, 2017.

[11] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1913–1924, 2011.

[12] V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Colleen, J. Yuan, and M. Liberman, "Articulatory trajectories for large-vocabulary speech recognition," in *Proc. IEEE ICASSP*, 2013, pp. 7145–7149.

[13] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, and E. Saltzman, "Articulatory features from deep neural networks and their role in speech recognition," in *Proc. IEEE ICASSP*, 2014, pp. 3017–3021.

[14] J. Yu, R. Su, L. Wang, and W. Zhou, "A multi-channel/multi-speaker interactive 3D audio-visual speech corpus in Mandarin," in *Proc. IEEE ISCSLP*, 2016, pp. 1–5.

[15] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[16] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. K. Kakumanu, and O. N. Garcia, "Audio/visual mapping with cross-modal hidden markov models," *IEEE Transactions on Multimedia*, vol. 7, no. 2, pp. 243–252, 2005.

[17] L. Xie and Z.-Q. Liu, "A coupled HMM approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, no. 8, pp. 2325–2340, 2007.

[18] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, no. 1, pp. 90–102, 1995.

[19] G. Takács, "Direct, modular and hybrid audio to visual speech conversion methods - a comparative study," in *Proc. ISCA INTERSPEECH*, 2009, pp. 2267–2270.

[20] S. Taylor, A. Kato, B. Milner, and I. Matthews, "Audio-to-visual speech conversion using deep neural networks," in *Proc. ISCA INTERSPEECH*, 2016, pp. 1482–1486.

[21] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. Thesis, The University of Edinburgh, UK., 2002.

[22] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE ICASSP*, 2013, pp. 6645–6649.

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

[24] X. Xie, X. Liu, and L. Wang, "Deep neural network based acoustic-to-articulatory inversion using phone sequence information," in *Proc. ISCA INTERSPEECH*, 2016, pp. 1497–1501.

[25] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. ISCA INTERSPEECH*, 2014, pp. 338–342.

[26] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[27] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in *Proc. NIPS*, 2013, pp. 190–198.

[28] P. J. Bell, M. J. F. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. C. Woodland, "Transcription of multi-genre media archives using out-of-domain data," in *Proc. IEEE SLT*, 2012, pp. 324–329.

[29] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Proc. IEEE ICASSP*, 2013, pp. 6975–6979.

[30] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Proc. ISCA INTERSPEECH*, 2013, pp. 3366–3370.

[31] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural network concepts to hybrid NN-HMM model for speech recognition," in *Proc. IEEE ICASSP*, 2012, pp. 4277–4280.

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

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.

[34] 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 *Proc. IEEE ICASSP*, 2015, pp. 4460–4464.