Exploiting Cross-Domain Visual Feature Generation for Disordered Speech Recognition

Shansong Liu[1*], *Xurong Xie*[1,2*], *Jianwei Yu*[1], *Shoukang Hu*[1], *Mengzhe Geng*[1], *Rongfeng Su*[2], *Shi-Xiong Zhang*[3], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3]Tencent AI Lab, Tencent

{ssliu,jwyu,skhu,mzgeng,xyliu,hmmeng}@se.cuhk.edu.hk, xrxie@ee.cuhk.edu.hk, rf.su@siat.ac.cn, auszhang@tencent.com

Abstract

Audio-visual speech recognition (AVSR) technologies have been successfully applied to a wide range of tasks. When developing AVSR systems for disordered speech characterized by severe degradation of voice quality and large mismatch against normal, it is difficult to record large amounts of high quality audio-visual data. In order to address this issue, a cross-domain visual feature generation approach is proposed in this paper. Audio-visual inversion DNN system constructed using widely available out-of-domain audio-visual data was used to generate visual features for disordered speakers for whom video data is either very limited or unavailable. Experiments conducted on the UASpeech corpus suggest that the proposed cross-domain visual feature generation based AVSR system consistently outperformed the baseline ASR system and AVSR system using original visual features. An overall word error rate reduction of 3.6% absolute (14% relative) was obtained over the previously published best system on the 8 UASpeech dysarthric speakers with audio-visual data of the same task.

Index Terms: Speech Disorders, Audio-Visual Speech Recognition, Audio-Visual Inversion, Cross-Domain Adaptation

1. Introduction

Human speech perception is inherently a bimodal process that uses both acoustic and visual information. Previous researches have shown that incorporating visual modality can improve the performance of speech recognition systems when being used in noisy environment [1, 2, 3, 4] or on impaired speech [5, 6, 7, 8, 9]. This motivates the use of audio-visual speech recognition (AVSR) systems for a wide range of applications targeting normal speech [10, 11, 12, 13, 14, 15].

The application of AVSR technologies to disordered speech faces two major challenges. First, it is difficult to record large amounts of high quality audio-visual (AV) data which is essential for developing AVSR systems using state-of-the-art data intensive deep learning techniques. Second, the visual information may not always be available, e.g. the commonly used benchmark UASpeech disordered speech corpus [16] only contains 8 speakers out of a total 16 with AV data. In addition, the visual data quality may be poorly usable due to the difficulty in tracking the lip regions of disordered speakers, resulting from their head movements or different angles facing the camera caused by severe medical conditions such as cerebral palsy and Parkinson disease.

*Equal contribution. Part of this work was done while the first author was an intern at Tencent AI lab.

Among the previous AVSR research for disordered speech [5, 6, 7, 8], these two issues remain largely unaddressed to date. Our previous work [8] attempted to address the first issue using Bayesian gated neural networks (BGNN) to more robustly use limited AV data containing severely impaired speakers with poor quality video data. However, the BGNN system is not applicable to speakers with no visual information. To the best of our knowledge, there was no previous work attempting to solve this missing visual modality problem for disordered speech recognition. More importantly, the overall quantity of the AV training data remains limited. This creates difficulty in developing larger and more powerful AVSR systems to further improve disordered speech recognition performance.

A possible solution to deal with the second issue on missing visual modality is to use the AV inversion approaches to generate missing visual features, inspired by early research on acoustic-to-articulatory [17, 18, 19, 20] and audio-visual inversion techniques [21, 22, 23]. However, a direct application of these methods is problematic. The existing AV disordered speech corpora are usually quite small in size and insufficient for AV inversion model training. Alternatively, more widely available, out-of-domain AV normal speech data, e.g. TV broadcast materials in the lip reading sentences 2 (LRS2) dataset [24], can be used to train AV inversion models. Unfortunately, this method cannot be directly applied to disordered speech given the large mismatch against normal speech, thus rendering the generated visual features unreliable for system development. The domain mismatch between normal and disordered speech needs to be minimized before such inversion systems trained on out-of-domain data can be used, as found in previous research [22] on cross-domain visual feature generation for domain mismatched normal speech data, e.g. wide band broadcast speech versus narrow band telephone conversation.

In order to address both data sparsity and missing visual modality issues mentioned above, a cross-domain visual feature generation approach was adopted. A high quality AV parallel dataset, i.e. the LRS2 dataset [24], was used to build deep AV inversion systems to generate visual features for 16 disordered speakers in the UASpeech corpus [16]. Cross-domain adaptation was performed to minimize the mismatch between the LRS2 and UASpeech audio data. The resulting cross-domain adapted AV inversion system was further applied to augmented disordered speech audio data. As a result, the total amount of AV data for AVSR system development was increased by up to nine folds compared to the 8 UASpeech speakers' AV subset.

The main contributions are summarized below. To the

best of our knowledge, this is the first work to explore cross-domain visual feature generation approaches for audio-visual disordered speech recognition. In contrast to previous research where both data sparsity and missing visual modality issues were unaddressed, this paper presents a dual purpose solution targeting both issues. Experiments conducted on the UASpeech corpus suggest that the proposed cross-domain visual feature generation based AVSR system consistently outperformed the baseline ASR system and AVSR system using original visual features. An overall word error rate reduction of 3.6% absolute (14% relative) was obtained over the previously published best system [8] on the 8 UASpeech dysarthric speakers with audio-visual data of the same task.

The rest of the paper is organized as follows. The baseline ASR and AVSR systems are described in section 2. Section 3 describes the AV inversion systems trained on the LRS2 dataset. Section 4 details the cross-domain visual feature generation approach. Experiments setup and results are shown in section 5. The last section concludes and discusses possible future work.

2. Audio-visual Speech Recognition

This section describes the DNN based ASR and AVSR systems architecture on which the experiments were conducted in this paper. The learning hidden unit contributions (LHUC) [25] based speaker adaptive training (SAT) was also used.

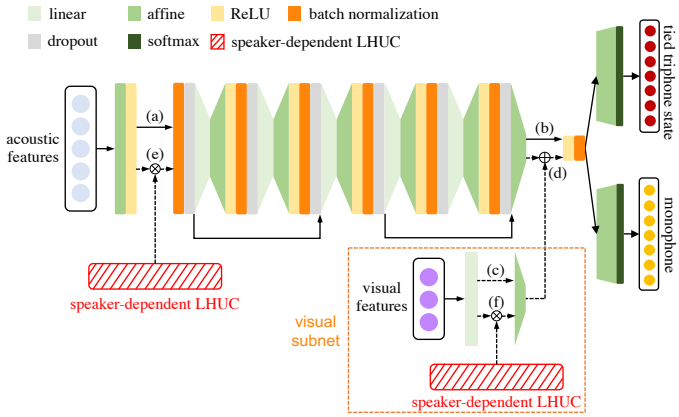

Figure 1: *The ASR and AVSR systems architecture developed for the UASpeech disordered speech corpus. Different connection combinations in this figure form different systems. For example, retaining connections (a) and (b) while discarding others leads to the ASR baseline system; Disconnecting (b) and retaining (a), (c) and (d) (for AV modality fusion) produces the AVSR baseline system. The connections (e) and (e)+(f) are used for LHUC speaker adaptive training of ASR and AVSR systems.*

The baseline ASR and AVSR systems share the same main structure regardless of the connections (c), (d), (e) and (f), shown in Fig. 1. The main structure is composed of seven hidden layers. Each hidden layer contains a set of neural operations performed in sequence: affine transformation (in green), rectified linear unit (ReLU) activation (in yellow) and batch normalization (in orange). To reduce network parameters, linear bottleneck projection layers (in light green) are applied to the inputs of the intermediate five hidden layers. The first six hidden layers are equipped with dropout operations (in grey) to avoid over-fitting. Softmax activation functions (in dark green) are used in the output layer. Additionally, we place two skip connections in the main structure to speed up the training process and circumvent the vanishing gradient problem. One skip connection is positioned between the outputs of the first and third layer, and the other is between the outputs of the fourth and sixth layer. The AVSR system is produced by adding a visual subnet using connections (c) and (d). The output of the visual subnet is then added to the sixth layer's output before the next ReLU activation.

Multi-task learning (MTL) [26] was adopted to train the systems illustrated in Fig. 1. The output targets for the two tasks are frame-level tied triphone states and monophone alignments respectively, obtained from a GMM-HMM system implemented using the HTK toolkit [27]. Incorporating monophone alignments can reduce the risk of over-fitting to unreliable triphone states computed from disordered speech. The loss function of the mutli-task learning is as follows:

$$\mathcal{L}_{MTL} = \lambda \cdot \mathcal{L}_{tristate} + (1-\lambda) \cdot \mathcal{L}_{mono} \quad (1)$$

The above loss function uses the cross-entropy criterion, where $\mathcal{L}_{tristate}$ is the loss of the tied triphone state task and $\mathcal{L}_{mono}$ is the loss of the monophone task. $0 \leq \lambda \leq 1$ is the weight ratio.

To handle the large variability among different dysarthric speakers, LHUC-SAT was used (see connections (e) and (f) for LHUC adaptation to the main architecture and visual subnet respectively in Fig. 1). During training, supervised estimation of the LHUC scaling factors was performed for each speaker. During test adaptation, an unsupervised LHUC adaptation was used, where the LHUC scaling factors were adapted.

3. Audio-visual Inversion

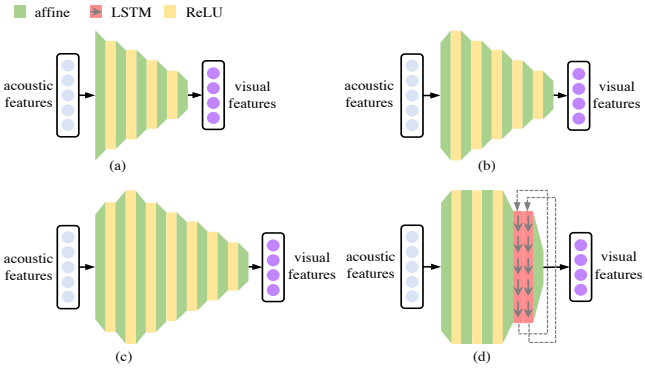

Figure 2: *Various AV inversion architectures investigated in this paper. (a), (b), (c) are DNN based AV inversion models, (d) is the LSTM based AV inversion model. The network inputs are splicing windowed frames of acoustic features, while the output is a single frame of visual features.*

In order to address the missing visual modality issue, an audio-visual inversion technique was employed. The objective of audio-visual inversion is to learn the mapping between acoustic and visual domain, hence a suitable inversion model is required. In previous research, Taylor et al. [21] explored the deep neural network (DNN) based AV inversion architectures, while Su et al. [22] adopted the long short term memory (LSTM) network with fully connected layers to train the AV inversion model. In this paper, we used the widely available, out-of-domain AV normal speech data, the LRS2 dataset [24], and investigated both inversion techniques, i.e. DNN and LSTM, to select a suitable AV inversion model by evaluating the AVSR performance on the LRS2 AV test set in comparison against a baseline AVSR

system using the original visual features only. The generated visual features were used in both AVSR system training and evaluation stages.

Table 1: *AVSR systems performance comparison (in WER) on the LRS2 AV test set. The visual features used in the AVSR systems are original visual features and generated visual features produced by the four investigated AV inversion systems.*

AVSR WER%				
original visual features	generated visual features			
	DNN			LSTM
	(a)	(b)	(c)	(d)
9.05	9.25	**8.92**	9.20	9.28

Various forms of AV inversion models are illustrated in Fig. 2. The first three (a), (b) and (c) adopted the feedforward DNN structure with different number of hidden layers, while (d) used additional LSTM layers following several fully connected layers. ReLU activation functions were applied to all models. The AVSR performance using generated visual features of the above four inversion models against that using original visual features are shown in Table 1. This table demonstrates that the AVSR systems using generated visual features from the four AV inversion models have comparable WER performance compared to that using original visual features. The DNN based AV inversion model (b) in Fig. 2 was selected for the rest of the paper with necessary cross adaptation being used considering its best performance among the four inversion models.

4. Cross-domain Visual Feature Generation

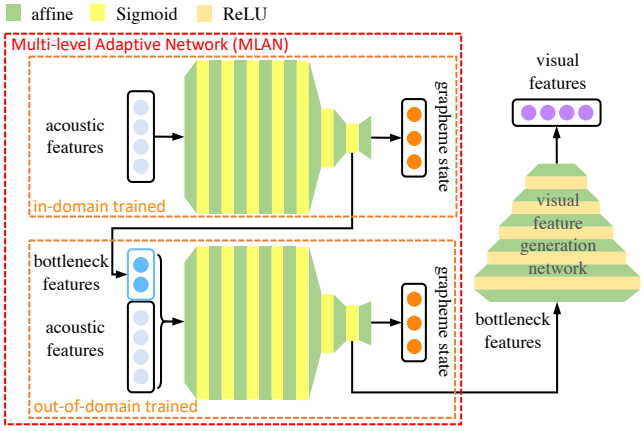

Figure 3: *Cross-domain visual feature generation system architecture. The left part is the MLAN network consisting of two DNN components, while the DNN in the right is the AV inversion model using the bottleneck features as the adapted "acoustic" inputs from the second DNN component of the MLAN network.*

The inversion systems (described in section 3) trained on the LRS2 AV normal speech data cannot be directly applied to disordered speech given the large mismatch against normal speech. This mismatch may lead the generated visual features unreliable to use for AVSR system development. There are multiple solutions can be applied to handle such mismatch, e.g. domain-adversarial neural networks (DANN) [28] or multi-level adaptive networks (MLAN) [29], which have been investigated by previous work [22] on normal speech datasets. Following [22], the MLAN based method was adopted to minimize the domain mismatch between normal and disordered speech in this paper.

As shown in the left of Fig. 3, the MLAN network consists of two DNN components, each with a bottleneck layer immediately before the output layer. The training process of such adaptive network is detailed as follows: the first-level DNN was trained with the audio data from the in-domain UASpeech corpus; the trained in-domain DNN was then used to produce bottleneck features for the out-of-domain audio data of the LRS2 AV normal speech; the next step was to train the second-level DNN using the out-of-domain LRS2 audio data concatenated with the bottleneck features computed from the previous step.

In this MLAN network during training stage, the out-of-domain LRS2 audio data was transformed to "UASpeech-like" data, via the bottleneck features produced by the first-level in-domain UASpeech audio data trained DNN. The UASpeech domain information inside the bottleneck features was removed by the second-level DNN component of the MLAN network, where the "UASpeech-like" LRS2 bottleneck features were reversed back to the LRS2 data domain. Then the bottleneck outputs of the second-level DNN component can be used as cross-domain adapted "acoustic" features for the following AV inversion model training (in the right of Fig. 3). When applying the MLAN network, we feedforward the UASpeech data through both components to generate "LRS2-like acoustic" features.

5. Experiments

5.1. Task Description and Experimental Setup

The UASpeech [16] is an isolated word recognition task containing 16 dysarthric and 13 control speakers, where only 8 dysarthric speakers have AV parallel data. The data was split into three blocks, B1, B2 and B3. The B1 and B3 of the 29 speakers (∼30.6h) were used for baseline ASR and AVSR systems' training. The performance evaluation was conducted on the B2 of the 16 dysarthric speakers (∼9h). The augmented UASpeech audio data (∼99.5h) for addressing the data sparsity issue were provided by the authors of [30].

The LRS2 dataset is one of the largest widely available AV normal speech datasets [24]. The train+validation and test sets of the LRS2 dataset were used for inversion model development in this paper. The acoustic features are 40-dimension mfcc features following [14]. The preprocessing step to obtain the original visual features of LRS2 video data follows our previous work [8] and the AVSR system description is detailed in [14].

In our ASR and AVSR experiments on UASpeech data, a 9-frame context window was used. The acoustic features are 80-dimension filter bank (FBK)+Δ features. The original visual features were the same with that used in our previous work [8], while the generated visual features were produced by the approach described in section 4. All the visual features used are 25 dimensions following our previous work [8]. In the main structure of both ASR and AVSR systems (in Fig. 1), the first six hidden layers contain 2000 neurons each, followed by dropout operations with a 20% dropout rate. The bottleneck dimension of the intermediate five hidden layers is 200. The seventh hidden layer contains 100 neurons. All the systems were trained by back-propagation based on RMSProp without pre-training. The weight ratio λ of the multi-task learning was set as 0.5. A uniform language model was used in decoding, following [31].

Table 2: *Comparison of the WER results produced by various ASR and AVSR systems investigated in this paper on the 16 UASpeech dysarthric speakers test set. The dysarthric speakers are grouped by their intelligibility levels, which are "Very low", "Low", "Mild" and "High". Data Aug. is the abbreviation of data augmentation.*

Systems		LHUC SAT	Data Aug.	WER%				
				Very low	Low	Mild	High	Average
1	audio-only	✗	✗	69.82	32.61	24.53	10.40	31.45
2	audio+original visual	✗	✗	69.70	32.59	24.35	9.67	31.14
3	audio+UA-syn visual	✗	✗	69.73	33.03	24.20	9.59	31.21
4	audio+LRS2-cross visual	✗	✗	**67.45**	**31.53**	**23.23**	10.28	**30.38**
5	audio-only	✓	✗	64.39	29.88	20.27	8.95	28.29
6	audio+original visual	✓	✗	63.80	29.99	20.35	8.66	28.11
7	audio+UA-syn visual	✓	✗	65.58	28.79	19.80	8.72	28.09
8	audio+LRS2-cross visual	✓	✗	**63.65**	29.32	**19.45**	9.20	**27.92**
9	audio-only	✗	✓	66.45	28.95	20.37	9.62	28.73
10	audio+LRS2-cross visual	✗	✓	**66.01**	29.22	20.59	9.67	28.76
11	audio-only	✓	✓	62.50	27.26	18.41	8.04	26.55
12	audio+LRS2-cross visual	✓	✓	**61.34**	27.90	**18.29**	9.22	26.84

5.2. In-domain AVSR

Prior to the proposed cross-domain visual feature generation approach (we use "Xdomain-AV approach" for simplification in the following), the in-domain UASpeech AV parallel data trained AV inversion model was investigated in this paper. The inversion model architecture is the same with the DNN based model (b) in Fig. 2. The training of this in-domain inversion model was based on the 8 dysarthric speakers' AV parallel data.

The performance of three in-domain systems is compared in this subsection (see Table 2, systems 1-3). System 1 is the in-domain ASR baseline. System 2 is the AVSR baseline using the UASpeech original visual features following our previous work [8]. Zeros were concatenated with the acoustic features for those speakers whose video data were missing. The AVSR system 3 used the generated visual features derived from the in-domain UASpeech AV inversion model for both training and test data. Comparing the results of these three lines we observe that no significant improvement can be obtained over the ASR baseline using either original visual features or in-domain inversion generated visual features.

5.3. Cross-domain Adapted AVSR

The system 4 in Table 2 is the AVSR system using the visual features generated by the Xdomain-AV approach. Same as the AVSR system 3, the generated visual features used in system 4 are applied to both training and test data. The line of system 4 indicates that the AVSR system using the generated visual features produced by the Xdomain-AV approach consistently outperforms systems 1-3 on the "Very low", "Low" and "Mild" dysarthric speaker groups, yet almost no improvement on the "High" group. The average WER reduction of system 4 compared to the ASR baseline is 1.07% absolute (3.4% relative).

LHUC based speaker adaptive training was further applied to model the variability among dysarthric speakers (see systems 5-8 in Table 2). Comparing the results, the similar trend is still observed. The Xdomain-AV AVSR system 8 produces the lowest average WER among the systems 5-8, while having a consistent WER reduction on the "Very low" and "Mild" groups.

In order to address the AV data sparsity issue, we acquired ∼99.5h augmented UASpeech audio data from [30]. Visual features were generated for the augmented audio data using our Xdomain-AV approach. Then the augmented audio data concatenated with generated visual features were added to the training set for systems 10,12. Consistent WER reductions on the "Very low" group over audio-only systems 9,11 with data augmentation were obtained using Xdomain-AV AVSR systems with or without LHUC speaker adaptive training. No further improvements could be obtained on other intelligibility levels.

5.4. Comparison with Previous Reported System

Table 3: *A comparison between the best WER result in this paper and published WER results on the 8 UASpeech dysarthric speakers with audio-visual data available.*

Systems	Avg WER%
Sheffield-2015 SAT based ASR [32]	33.1
CUHK-2018 DNN ASR [33]	30.2
CUHK-2019 SD BGNN AVSR [8]	25.7
Xdomain-AV AVSR in this paper	**22.1**

We compare the best average WER result of the AVSR system using the visual features generated by the Xdomain-AV approach with previously published best ASR and AVSR systems available on the 8 UASpeech dysarthric speakers (see in Table 3). We can observe that our AVSR system using the generated visual features produced by the Xdomain-AV approach achieves the lowest WER and gives a 3.6% absolute (14% relative) WER reduction compared with our previous work [8].

6. Conclusion

In this paper, we present the first work to explore cross-domain visual feature generation approaches for audio-visual disordered speech recognition. The experiments conducted indicate that this method is useful for disordered speech recognition where the audio-visual data is often very limited. Our future research will focus on improving audio-visual data generation and augmentation techniques.

7. Acknowledgement

This research is supported by Hong Kong Research Grants Council General Research Fund No.14200218, Theme Based Research Scheme T45-407/19N, Shun Hing Institute of Advanced Engineering Project No. MMT-p1-19 and ShenZhen Fundamental Research Program KQJSCX20170731163308665. We thank Dr. Feifei Xiong for insightful discussions.

8. References

[1] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7596–7599.

[2] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2130–2134.

[3] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[4] S. Zhang, M. Lei, B. Ma, and L. Xie, "Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6570–6574.

[5] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal speech recognition of a person with articulation disorders using AAM and MAF," in *2010 IEEE International Workshop on Multimedia Signal Processing*. IEEE, 2010, pp. 517–520.

[6] A. Farag, M. E. Adawy, and A. Ismail, "A robust speech disorders correction system for arabic language using visual speech recognition." 2013.

[7] E. S. Salama, R. A. El-Khoribi, and M. E. Shoman, "Audio-visual speech recognition for people with speech disorders," *International Journal of Computer Applications*, vol. 96, no. 2, 2014.

[8] S. Liu, S. Hu, Y. Wang, J. Yu, R. Su, X. Liu, and H. Meng, "Exploiting visual features using bayesian gated neural networks for disordered speech recognition," *Proc. Interspeech 2019*, pp. 4120–4124, 2019.

[9] S. Hu, S. Liu, H. F. Chang, M. Geng, J. Chen, T. K. H. Chung, J. Yu, K. H. Wong, X. Liu, and H. Meng, "The CUHK Dysarthric Speech Recognition Systems for English and Cantonese," *Proc. Interspeech 2019*, pp. 3669–3670, 2019.

[10] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[11] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2002, no. 11, p. 783042, 2002.

[12] R. Su, L. Wang, and X. Liu, "Multimodal learning using 3D audio-visual data for audio-visual speech recognition," in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 40–43.

[13] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1290–1302, 2018.

[14] J. Yu, S. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, and D. Yu, "Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6984–6988.

[15] J. Yu, B. Wu, R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, D. Yu, X. Liu, and H. Meng, "Audio-visual multi-channel recognition of overlapped speech," *Interspeech*, 2020.

[16] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[17] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4450–4454.

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

[19] V. Mitra, G. Sivaraman, C. Bartels, H. Nam, W. Wang, C. Espy-Wilson, D. Vergyri, and H. Franco, "Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5205–5209.

[20] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, "Dnn-based acoustic-to-articulatory inversion using ultrasound tongue imaging," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[21] S. Taylor, A. Kato, B. Milner, and I. Matthews, "Audio-to-visual speech conversion using deep neural networks," 2016.

[22] R. Su, X. Liu, L. Wang, and J. Yang, "Cross-domain deep visual feature generation for mandarin audio–visual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 185–197, 2019.

[23] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *AAAI*, vol. 33, 2019, pp. 9299–9306.

[24] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[25] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1450–1463, 2016.

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

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book," *Cambridge university engineering department*, vol. 3, p. 75, 2006.

[28] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[29] P. Bell, M. 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 *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 324–329.

[30] M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu, and H. Meng, "Investigation of Data Augmentation Techniques for Disordered Speech Recognition," in *INTERSPEECH*, 2020.

[31] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[32] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 65–71.

[33] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Lam, X. Wu, K. H. Wong, X. Liu, and H. Meng, "Development of the CUHK Dysarthric Speech Recognition System for the UA Speech Corpus," in *Interspeech*, 2018, pp. 2938–2942.