INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Exploiting Visual Features using Bayesian Gated Neural Networks for Disordered Speech Recognition

Shansong Liu[1], *Shoukang Hu*[1], *Yi Wang*[2], *Jianwei Yu*[1], *Rongfeng Su*[3], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]University of Cambridge, Cambridge, UK
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

{ssliu,skhu,jwyu,xyliu,hmmeng}@se.cuhk.edu.hk, yw454@cam.ac.uk, rf.su@siat.ac.cn

Abstract

Automatic speech recognition (ASR) for disordered speech is a challenging task. People with speech disorders such as dysarthria often have physical disabilities, leading to severe degradation of speech quality, highly variable voice characteristics and large mismatch against normal speech. It is also difficult to record large amounts of high quality audio-visual data for developing audio-visual speech recognition (AVSR) systems. To address these issues, a novel Bayesian gated neural network (BGNN) based AVSR approach is proposed. Speaker level Bayesian gated control of contributions from visual features allows a more robust fusion of audio and video modality. A posterior distribution over the gating parameters is used to model their uncertainty given limited and variable disordered speech data. Experiments conducted on the UASpeech dysarthric speech corpus suggest the proposed BGNN AVSR system consistently outperforms state-of-the-art deep neural network (DNN) baseline ASR and AVSR systems by 4.5% and 4.7% absolute (14.9% and 15.5% relative) in word error rate.
Index Terms: Speech Disorder, Audio-Visual Speech Recognition, Bayesian Gated Neural Network

1. Introduction

Speech disorders lead to the disruption of normal speech. They affect millions of people worldwide and the quality of their life. Dysarthria is a common form of speech disorders associated with neuromotor conditions [1], such as Parkinson disease and cerebral palsy [2, 3], as well as brain damages due to stroke or head injuries. Speech disorders lead to severe degradation of speech quality, highly variable voice characteristics and large mismatch against normal speech. In addition, it is difficult to collect high quality speech data in large quantities for automatic speech recognition (ASR) systems development [4]. For the above reasons, disordered speech recognition is a very challenging research problem to date [5, 6, 7, 8, 9].

Human speech generation is inherently a bimodal process based on audio-visual representation. This is also true for speech perception. The visual information is invariant to acoustic signal corruption and can provide complementary information to the speech recognizer. This motivates the use of visual information to improve speech recognition performance by developing audio-visual speech recognition (AVSR) systems [10, 11, 12, 13, 14, 15, 16]. Visual information used in human speech perception mainly constitutes lip motion, head movement, facial expression and body gesture. Among these, lip information is the primary form of visual information that is incorporated in current AVSR systems [17, 18, 19, 20].

Earlier forms of AVSR approaches were based on hidden Markov models (HMMs), such as multi-stream HMMs [10], product HMMs [11], coupled HMMs [12] and factorial HMMs [21]. In recent years, deep learning has been widely adopted in AVSR systems development [13, 22, 23, 24, 25]. For example, Huang et al. [13] showed that deep belief networks reduced the word error rate (WER) by 21% relative over HMM models on the 5.3 hours in-house collected audio-visual data. Ninomiya et al. [22] used deep nerual network (DNN) bottleneck features based on both audio and video inputs on the 2.6 hours CENSREC-1-AV dataset [26] and achieved from 52% to 69% relative WER reduction over the HMM baseline systems.

However, most previous AVSR research were conducted to develop AVSR systems for normal speech data. Only a few AVSR systems were constructed for speakers with speech disorders [27, 28, 29, 30, 31]. The majority of these previous studies used HMM based AVSR system architecture. The research presented in [29] used the audio-visual data from 10 speakers of the UASpeech corpus [32], the largest disordered speech corpus available to date. A WER of 39% was obtained. In contrast, state-of-the-art DNN based ASR systems developed in [9] produced a much lower WER of 23% on the same set of speakers. This suggests that the development of AVSR technologies for disordered speech still falls behind that of ASR systems and there is a pressing need to improve AVSR technologies for such data to improve the quality of life of those affected.

In addition to the well known degradation of voice quality, there are several new challenges when developing AVSR systems for people with speech disorders. First, their underlying medical conditions such as cerebral palsy and Parkinson disease combined with possibly co-occurring disabilities increase the difficulty to record high quality visual data. For example, head movements and different angles facing the camera are often found. These make the accurate detection of lip regions very difficult, and the subsequent extracted visual features unreliable to use. Second, in common with the audio data, the diverse causes leading to speech disorders and the resulting symptoms create a large variability among individual impaired speakers. Finally, it is generally difficult to collect large amounts of audio-visual data from people suffering from speech impairment.

In order to address these issues, we propose a novel Bayesian gated neural network (BGNN) based AVSR architecture in this paper. This is realized by positioning an additional multiplicative gating layer [20] between the input and first hidden layer. This layer's outputs are used to dynamically weight the contributions from visual features before they are further concatenated with acoustic features. This allows a more flexible fusion of acoustic and visual features that can learn to suppress non-discriminant visual data. It is generally possible to add gates to both modalities. However, our scope in this pa-

http://dx.doi.org/10.21437/Interspeech.2019-1536

per is to investigate the integration and selection on visual data. Speaker dependent BGNN models are constructed to handle speaker level variability. In order to address the data sparsity issue, a posterior distribution over the gating layer weight and bias parameters is used to model their uncertainty given limited and variable data. An efficient variational inference based approach [33] is also used in BGNN system training.

The main contributions of this paper are summarized below. First, to the best of our knowledge, this paper proposes the first use of Bayesian gated neural networks for AVSR systems, in contrast to previous use of simple concatenation of acoustic and visual features [24], and the conventional, non-Bayesian gated neural networks [20] using fixed point parameter estimation. Second, this paper presents the first use of deep learning based AVSR approaches for disordered speech on the largest available UASpeech corpus. Finally, the proposed BGNN AVSR system outperforms both the previously published best DNN ASR system in [9], and the baseline DNN AVSR system constructed using feature concatenation by 4.5% and 4.7% absolute (14.9% and 15.5% relative) in WER. Consistent improvements were also obtained over the conventional gated neural network (GNN) based AVSR system.

The paper is structured as follows. The basic DNN AVSR system architecture is described in section 2. The proposed BGNN model is presented in section 3. Section 4 elaborates the experiments and results. The last section concludes and discusses possible future work.

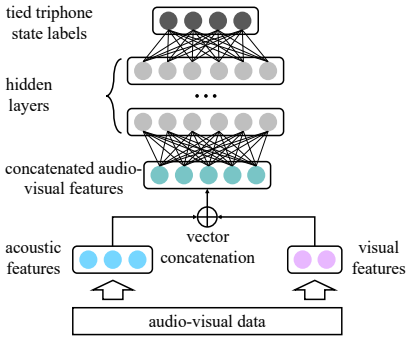

Figure 1: *The framework of the standard DNN based AVSR architecture. Acoustic and visual features are concatenated at the input layer before they are fed into the subsequent hidden layers. Network outputs are the tied triphone state labels.*

2. Audio-visual Speech Recognition

A commonly used approach in DNN based AVSR systems is to concatenate the acoustic and visual features at the input layer [23, 24, 25], as shown in the example of Fig. 1. Given an input vector $\mathbf{z}_t^{(l-1)}$ from $(l-1)$-th layer at t-th frame, a standard DNN AVSR system computes the output $h_i^{(l)}(\mathbf{z}_t^{(l-1)})$ of the i-th node in the l-th layer using Eqn. (1).

$$h_i^{(l)}(\mathbf{z}_t^{(l-1)}) = \phi\left(\boldsymbol{\theta}_i^{(l)} \bullet \mathbf{z}_t^{(l-1)}\right) \tag{1}$$

where $\mathbf{z}_t^{(l-1)} = \left[h_1^{(l-1)}(\mathbf{z}_t^{(l-2)}), \cdots, h_d^{(l-1)}(\mathbf{z}_t^{(l-2)}), 1\right]$ is the input vector fed into the l-th hidden layer, $\boldsymbol{\theta}_i^{(l)} = \left[w_{i,1}^{(l)}, \cdots, w_{i,d}^{(l)}, b^{(l)}\right]$ denotes the node's weight vector, $\phi(\cdot)$ is the activation function, and $\bullet$ denotes the dot product.

Acoustic and visual features are concatenated at the input layer, where $\mathbf{z}_t^{(0)} = [\mathbf{x}_t^a \oplus \mathbf{x}_t^v, 1]$. $\mathbf{x}_t^a$ and $\mathbf{x}_t^v$ are the acoustic and visual feature vectors of t-th frame, respectively. $\oplus$ denotes the vector concatenation operation. The output layer targets are tied triphone state labels.

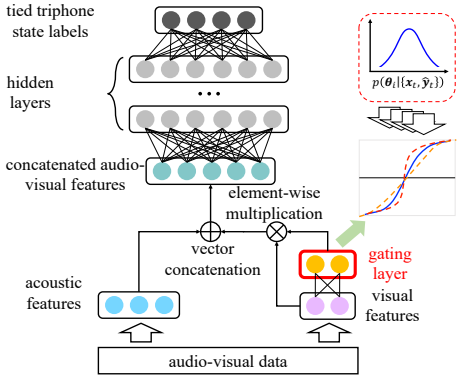

Figure 2: *A Bayesian gated DNN based AVSR system architecture. In contrast to the conventional gated DNN, a posterior distribution (top right corner) over the gating parameters is used to model the uncertainty given limited and variable visual data.*

3. Bayesian Gated Neural Network for Audio-visual Speech Recognition

The proposed Bayesian gated neural network (BGNN) model for AVSR system development is described in this section, as shown in Fig. 2. Compared to DNN AVSR system, a gating layer is placed at the input layer to dynamically weight the contributions from visual features, and a posterior distribution over the gating parameters is also applied to model the uncertainty given limited and variable disordered speech data.

The focus of this paper is the incorporation of selected visual features to help speech recognition, hence the gated control is only applied to the visual modality in our proposed model. The gated input layer outputs $\mathbf{z}_t^{(0)}$ are computed as:

$$\begin{aligned} \mathbf{z}_t^v &= [\mathbf{x}_t^v, 1] \\ h_i^{(0)}(\mathbf{z}_t^v) &= \phi\left(\boldsymbol{\theta}_i^{(0)} \bullet \mathbf{z}_t^v\right) \\ \mathbf{z}_t^{(0),v} &= \mathbf{x}_t^v \otimes h^{(0)}(\mathbf{z}_t^v) \\ \mathbf{z}_t^{(0)} &= [\mathbf{x}_t^a \oplus \mathbf{z}_t^{(0),v}, 1] \end{aligned} \tag{2}$$

where the gating layer is denoted as the 0-th hidden layer. $\otimes$ and $\oplus$ denote the element-wise multiplication and vector concatenation, respectively. The activation function $\phi(\cdot)$ is a sigmoid function, whose outputs vary between 0 and 1.

The general form of the hidden output with Bayesian learning [34] is as follows:

$$h_i^{(l)}(\mathbf{z}_t^{(l-1)}) = \int \phi\left(\boldsymbol{\theta}_i^{(l)} \bullet \mathbf{z}_t^{(l-1)}\right) p(\boldsymbol{\theta}_i^{(l)}) d\boldsymbol{\theta}_i^{(l)} \tag{3}$$

where $p(\boldsymbol{\theta}_i^{(l)}) = p(\boldsymbol{\theta}_i^{(l)} \mid \{\mathbf{x}_t, \widehat{\mathbf{y}}_t\})$ denotes the node dependent activation parameter posterior distribution to be learned from training data $\{\mathbf{x}_t, \widehat{\mathbf{y}}_t\}$ ($\mathbf{x}_t, \widehat{\mathbf{y}}_t$ are the input data and its corresponding triphone state label at t-th frame). In our scenario, we only perform Bayesian learning on the gating parameters, hence the Eqn. (3) can be rewritten as the specialized form Eqn. (4).

$$h_i^{(0)}(\mathbf{z}_t^v) = \int \phi\left(\boldsymbol{\theta}_i^{(0)} \bullet \mathbf{z}_t^v\right) p(\boldsymbol{\theta}_i^{(0)}) d\boldsymbol{\theta}_i^{(0)} \tag{4}$$

To estimate the hyper-parameters of the posterior distribution $p(\boldsymbol{\theta}_i^{(0)})$, the standard back-propagation algorithm needs to be modified to include two additional steps. First, to calculate the variational lower bound approximate over the integration of the model parameters. Second, a sampling step applied to the first term of the lower bound is required to obtain the gradient statistics for updating the hyper-paramters in standard back-propagation. These changes allow all layers including the gating layer of the network to be updated using back-propagation. In this paper, we use variational inference approach [33] to approximate the integration in Eqn. (4). For notation simplicity, we consider the parameters $\boldsymbol{\theta} = \boldsymbol{\theta}_i^{(0)}$ as the gating parameters at the i-th gating layer node. By applying Jensen's inequality, we calculate the evidence lower bound of the cross-entropy criterion, or equivalently the log-likelihood (see Eqn. (5)) of tied HMM state sequence Y given input acoustic feature vector sequence X, with visual features optionally appended.

$$\begin{aligned}\log P(\mathbf{Y} \mid \mathbf{X}) &= \log \int P(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X}) P_r(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\geq \underbrace{\int q(\boldsymbol{\theta}) \log P(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X}) d\boldsymbol{\theta}}_{\mathcal{L}_1} - \underbrace{KL(q(\boldsymbol{\theta}) \| P_r(\boldsymbol{\theta}))}_{\mathcal{L}_2} = \mathcal{L}\end{aligned} \tag{5}$$

where $P_r(\boldsymbol{\theta})$ denotes gating parameters prior distribution, $q(\boldsymbol{\theta})$ is the variational approximation of gating parameters posterior distribution $p(\boldsymbol{\theta})$. We assume that the variational distribution q and the prior distribution P_r are both Gaussian distributions, following [35], i.e. $P_r(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r^2)$, $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. $KL(q \| P_r)$ is the Kullback-Leibler (KL) divergence between q and P_r.

The first term $\mathcal{L}_1$ in Eqn. (5) can be efficiently approximated by Monte Carlo sampling method.

$$\begin{aligned}\mathcal{L}_1 &= \sum_{t=1}^{T} \int q(\boldsymbol{\theta}) \log P(\mathbf{y}_t \mid \boldsymbol{\theta}, \mathbf{x}_t) d\boldsymbol{\theta} \\ &\approx \frac{1}{N} \sum_{t=1}^{T} \sum_{k=1}^{N} \log P(\mathbf{y}_t \mid \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}_k, \mathbf{x}_t)\end{aligned} \tag{6}$$

where T is the total number of frames in the training data, $\boldsymbol{\epsilon}_k = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the k-th sample.

The KL divergence between q and P_r of the second term $\mathcal{L}_2$ in Eqn. (5) can be simplified as follows,

$$\mathcal{L}_2 = \sum_j \left\{ \log \frac{\sigma_{r,j}}{\sigma_j} + \frac{\sigma_j^2 + (\mu_j - \mu_{r,j})^2}{2\sigma_{r,j}^2} - \frac{1}{2} \right\} \tag{7}$$

where μ_j and σ_j are the j-th component of variational posterior distribution hyper-parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\mu_{r,j}$ and $\sigma_{r,j}$ are the j-th component of prior distribution hyper-parameters $\boldsymbol{\mu}_r$ and $\boldsymbol{\sigma}_r$. Then we can calculate the gradient statistics for the hyper-parameters $\lambda = \{\mu_j, \sigma_j\}$ as the following,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu_j} &= \frac{1}{N} \sum_{t,k=1}^{T,N} \frac{\partial \log P(\mathbf{y} \mid \mathbf{x}, \lambda, \boldsymbol{\epsilon}_k)}{\partial \mu_j} - \frac{T_b}{T} \frac{(\mu_j - \mu_{r,j})}{\sigma_j^2} \\ \frac{\partial \mathcal{L}}{\partial \sigma_j} &= \frac{1}{N} \sum_{t,k=1}^{T,N} \frac{\partial \log P(\mathbf{y} \mid \mathbf{x}, \lambda, \boldsymbol{\epsilon}_k)}{\partial \sigma_j} - \frac{T_b}{T} \left(\frac{\sigma_j^2 - \sigma_{r,j}^2}{\sigma_j \sigma_{r,j}^2} \right)\end{aligned} \tag{8}$$

where T_b is the number of frames in a minibatch [36]. Then the standard back-propagation method can be applied to the calculation of the two gradient terms $\frac{\partial \log P(\mathbf{y} \mid \mathbf{x}, \lambda, \boldsymbol{\epsilon}_k)}{\partial \mu_j}$ and $\frac{\partial \log P(\mathbf{y} \mid \mathbf{x}, \lambda, \boldsymbol{\epsilon}_k)}{\partial \sigma_j}$ for updating hyper-parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$. During model evaluation, the mean of the gating parameter $\boldsymbol{\mu}$ of the posterior distribution $p(\boldsymbol{\theta})$ is used as the drawn sample to compute the gating layer outputs.

(a) F04 (b) F05 (c) M11 (d) M12

Figure 3: *Example video snapshots of four UASpeech speakers F04, F05, M11 and M12 with different head movement patterns and various angles facing the camera.*

4. Experiments

4.1. Task Description and Experimental Setup

The UASpeech was recorded by an 8-channel microphone array and a video camera, including 7 channels with segmented single word audio segments and 1 channel with whole length unsegmented videos (having both audio and video streams) [32].

The UASpeech is an isolated word recognition task including 16 dysarthric speakers. 12 dysarthric speakers have both audios and videos, but only 8 out of the 12 speakers' videos were provided with video time segment labels (start and end time stamps of each word for chopping the whole length video into single word video segments). We utilized these 8 dysarthric speakers' audio-visual data to conduct our AVSR experiments. All speakers were required to repeat 455 distinct words. These words were distributed into three blocks. The block 1 (B1) and block 3 (B3) were treated as the training set, leaving the remaining block 2 (B2) as the test set.

The authors in [9] tried a range of deep learning based acoustic models including feed-forward DNNs, time delayed neural networks [37] and long short-term memory recurrent neural networks [38] on the UASpeech corpus. Among these, the feed forward DNN produced the lowest word error rate (WER). Therefore it is used as the ASR baseline system in this paper. All the neural network models developed in this paper are built by PyTorch [39].

In our experiments, a 9-frame context window was used in both ASR and AVSR systems' inputs. Acoustic features fed to neural networks are 80-dimension filter banks (FBKs)+Δ features. Target tied triphone state labels were produced by speaker dependent GMM-HMM models. Speaker dependent neural network acoustic models used 5 hidden layers of 500 neurons each, which is applicable for all experimented models in this paper. For the gated DNN and Bayesian gated DNN AVSR systems, an additional hidden layer having the same size with the input visual feature dimension was added. During training, we performed a layer-wise pretraining, then fine-tuned the whole network using SGD optimization method associated with a NEWBOB learning rate scheduler until no validation accuracy improvement was gained. For performance evaluation, the frame level output probability tables were fed to the HDecode in the HTK toolkit [40] to produce recognition outputs. A word grammar network was used in decoding, following [5].

Table 1: *Performance of baseline ASR systems v.s. AVSR, non-Bayesian GNN AVSR, and proposed BGNN AVSR systems with two types of visual features on the 8 UASpeech dysarthric speakers with audio-visual data available. All the systems are speaker dependent.*

Speaker ID	Intelligibility [5]	WER% (numbers in brackets indicate absolute WER reduction over ASR system)						
		ASR-DNN	AVSR (DCT-LDA)			AVSR (AE-LDA)		
			DNN	GNN	BGNN	DNN	GNN	BGNN
F02	low	49.6	49.6	48.1	46.8	46.4	45.4	**40.8** (-8.8)
F04	mid	29.2	34.8	32.6	27.6	32.9	29.9	**24.0** (-5.2)
F05	high	9.7	11.6	10.4	9.0	11.4	9.8	**8.0** (-1.7)
M08	high	11.6	16.1	13.3	11.1	11.4	10.4	**9.2** (-2.4)
M11	mid	32.3	38.6	35.3	31.8	37.4	35.2	**32.5** (+0.2)
M12	very low	65.7	62.7	62.2	60.5	59.5	58.4	**54.0** (-11.7)
M14	high	20.3	22.7	20.3	19.3	20.2	19.8	**15.6** (-4.7)
M16	low	29.5	29.5	29.0	28.0	29.2	27.1	**26.7** (-2.8)
Avearge		30.3	32.6	30.8	28.7	30.4	28.9	**25.7**

4.2. Audio and Video Data Preoprocessing

The whole length unsegmented videos were chopped into segmented word video segments. We slightly adjusted the time boundaries of the 7 channels' single word audio segments using the time boundaries of the segmented single word video segments. Excessive amounts of silence of the audio segments were removed by following the strategy described in [9].

For the visual feature extraction process, the video segments were first upsampled to match the frame numbers of acoustic features of each word. We employed an off-the-shelf face alignment network [41] to detect lip landmarks on the upsampled video data. From Fig. 3 we notice that the lips of the speakers are not in the horizontal view due to different head movements, so we applied affine transformation to detected lip regions to make them horizontal. Since lip regions are not the same size, we resized them to 128*128 pixels, following [29]. Afterwards, two unsupervised dimension reduction techniques were investigated respectively, i.e. discrete cosine transform (DCT) and autoencoder (AE), to downsize the lip regions to 40-dimension vectors. Finally, we applied linear discriminant analysis (LDA) to further reduce the size of visual feature vectors to 25 dimensions. The two types of dimension reducted visual features are denoted as DCT-LDA and AE-LDA.

4.3. Performance of Baseline ASR and AVSR Systems

The performance of two baseline systems are shown in Table 1 (2nd, 3rd and 6th columns). Our baseline ASR systems produced a competitive average WER (30.3%[1]) compared to [9] on the 8 dysarthric speakers. For baseline AVSR systems, the average WER degradation is observed on both two types of visual features, i.e. DCT-LDA and AE-LDA. This suggests that a seletion mechanism is required to find a more robust integration of acoustic and visual features.

4.4. Performance of GNN and BGNN AVSR Systems

The performance of the AVSR systems constructed using the proposed BGNN approach and the conventional non-Bayesian GNN model is shown in Table 1) (4th, 5th, 7th and 8th columns). There is an obvious trend that no matter which visual features are used, the proposed BGNN approach outperforms the baseline ASR and AVSR systems. However, using AE-LDA visual features with BGNN approach provides more significant average WER reduction, which are 4.5% and 4.7% absolute (14.9% and 15.5% relative), over baseline ASR and AVSR systems. Consistent improvements are also observed over the conventional non-Bayesian GNN AVSR systems.

[1]The average WER of DNN ASR systems [9] built by HTK on the 8 speakers is 30.2%. The authors of [9] provide the speaker level WERs.

4.5. BGNN AVSR Systems v.s. Published ASR Systems

We compare the best average WER result of the AVSR systems using the proposed BGNN AVSR architecture with previously published best ASR systems available on the 8 UASpeech dysarthric speakers (see in Table 2). This comparison shows that our proposed BGNN AVSR architecture using AE-LDA visual features (last row in Table 2) in this paper achieves the lowest WER.

Table 2: *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-2012 ASR [5]	39.5
Sheffield-2015 ASR [8]	33.1
CUHK-2018 ASR [9]	30.2[1]
BGNN AVSR	**25.7**

5. Conclusion

In this paper, we present the first work using Bayesian gated neural network for AVSR systems development. To the best of our knowledge, this paper also presents the first use of deep learning based AVSR approaches for disordered speech on the largest available dysarthric speech corpus–UASpeech. The proposed BGNN AVSR systems achieve the lowest word error rate compared to baseline ASR and AVSR systems in this paper, as well as those previously published best ASR systems. Possible future research will focus on improving speaker adaptation and adaptive training techniques to handle variability among dysarthric speakers.

6. Acknowledgement

This research is supported by Hong Kong Research Grants Council General Research Fund No. 14200218 and Shun Hing Institute of Advanced Engineering Project No. MMT-p1-19.

7. References

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, *Motor speech disorders*. Saunders, 1975.

[2] T. L. Whitehill and V. Ciocca, "Speech errors in cantonese speaking adults with cerebral palsy," *CLIN LINGUIST PHONET*, vol. 14, no. 2, pp. 111–130, 2000.

[3] S. Scott, F. Caird, and B. Williams, "Evidence for an apparent sensory speech disorder in parkinson's disease," *J NEUROL NEUROSUR PS*, vol. 47, no. 8, pp. 840–843, 1984.

[4] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *ASSIST TECHNOL*, vol. 22, no. 2, pp. 99–112, 2010.

[5] 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.

[6] H. Christensen, M. Aniol, P. Bell, and et al., "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *INTERSPEECH*, 2013.

[7] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *2014 12th International Conference on Signal Processing (ICSP)*. IEEE, 2014, pp. 505–509.

[8] 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.

[9] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Lam, X. Wu, and K. Ho, "Development of the cuhk dysarthric speech recognition system for the uaspeech corpus," *Proc. Interspeech 2018*, pp. 2938–2942, 2018.

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

[11] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 1–6.

[12] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled hmm for audio-visual speech recognition," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2002, pp. II–2013.

[13] 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.

[14] 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.

[15] 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.

[16] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *ICMI*, 2018.

[17] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE T MULTIMEDIA*, vol. 17, no. 5, pp. 603–615, 2015.

[18] A. Narwekar and P. K. Ghosh, "Prav: A phonetically rich audio visual corpus," *INTERSPEECH*, 2017.

[19] A. H. Abdelaziz, "Ntcd-timit: A new database and baseline for noise-robust audio-visual speech recognition," in *INTERSPEECH*, 2017.

[20] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 7, pp. 1286–1298, 2018.

[21] A. V. Nefian, L. Liang, X. Pi, and et al., "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP J ADV SIG PR*, vol. 2002, no. 11, p. 783042, 2002.

[22] H. Ninomiya, N. Kitaoka, S. Tamura, and et al., "Integration of deep bottleneck features for audio-visual speech recognition," in *INTERSPEECH*, 2015.

[23] K. Noda, Y. Yamaguchi, K. Nakadai, and et al., "Audio-visual speech recognition using deep learning," *APPL INTELL*, vol. 42, no. 4, pp. 722–737, 2015.

[24] Y. Miao and F. Metze, "Open-domain audio-visual speech recognition: A deep learning approach," in *INTERSPEECH*, 2016.

[25] T. Afouras, J. S. Chung, A. Senior, and et al., "Deep audio-visual speech recognition," *arXiv preprint arXiv:1809.02108*, 2018.

[26] S. Tamura, C. Miyajima, N. Kitaoka, and et al., "Censrec-1-av: An audio-visual corpus for noisy bimodal speech recognition," in *AVSP*, 2010.

[27] C. Miyamoto, Y. Komai, T. Takiguchi, and et al., "Multimodal speech recognition of a person with articulation disorders using aam and maf," in *MMSP*, 2010.

[28] A. Farag, M. El Adawy, and A. Ismail, "A robust speech disorders correction system for arabic language using visual speech recognition," *Biomedical Research*, vol. 24, no. 2, 2013.

[29] 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.

[30] Y. Takashima, Y. Kakihara, R. Aihara, and et al., "Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss," *IPSJ T CVA*, vol. 7, pp. 64–68, 2015.

[31] Y. Takashima, T. Takiguchi, Y. Ariki, and et al., "Audio-visual speech recognition for a person with severe hearing loss using deep canonical correlation analysis," *CHAT*, 2017.

[32] 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.

[33] A. Graves, "Practical variational inference for neural networks," in *ADV NEUR IN*, 2011, pp. 2348–2356.

[34] S. Hu, M. W. Lam, X. Xie, S. Liu, J. Yu, X. Wu, X. Liu, and H. Meng, "Bayesian and gaussian process neural networks for large vocabulary continuous speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6555–6559.

[35] D. Barber and C. M. Bishop, "Ensemble learning in bayesian neural networks," *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, vol. 168, pp. 215–238, 1998.

[36] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Advances in Neural Information Processing Systems*, 2015, pp. 2575–2583.

[37] A. Waibel, T. Hanazawa, G. Hinton, and et al., "Phoneme recognition using time-delay neural networks," in *Readings in speech recognition*, 1990, pp. 393–404.

[38] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[39] K. Nikhil, *Introduction to PyTorch*. Berkeley, CA: Apress, 2017, pp. 195–208.

[40] S. Young, G. Evermann, M. Gales, and et al., "The htk book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[41] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030.