ROBUST AUDIO-VISUAL SPEECH RECOGNITION USING BIMODAL DFSMN WITH MULTI-CONDITION TRAINING AND DROPOUT REGULARIZATION

Shiliang Zhang[1], *Ming Lei*[1], *Bin Ma*[1], *Lei Xie*[2]

[1]Machine Intelligence Technology, Alibaba Group
[2]School of Computer Science, Northwestern Polytechnical University
{sly.zsl, lm86501, b.ma}@alibaba-inc.com, lxie@nwpu.edu.cn

ABSTRACT

Audio-visual speech recognition (AVSR) is thought to be one of the potential solutions for robust speech recognition, especially in noisy environments. Compared to audio only speech recognition, the major issues of AVSR include the lack of publicly available audio-visual corpora and the need of robust knowledge fusion of both speech and vision. In this work, based on the recently released NTCD-TIMIT audio-visual corpus, we address the challenges of AVSR through three aspects: 1) optimal integration of acoustic and visual information; 2) robust performance with multi-condition training; 3) robust modeling against missing visual information during decoding. We propose a bimodal-DFSMN to jointly learn feature fusion and acoustic modeling, and utilize a per-frame dropout approach to enhance the robustness of AVSR system against the missing of visual modality. In the experiments, we construct two setups based on the NTCD-TIMIT corpus that consists of 5 hours clean training data and 150 hours multi-condition training data, respectively. As a result, we achieve a phone error rate of 12.6% on clean test set and an average phone error rate of 26.2% on all test sets (clean, various SNRs, various noise types), which both dramatically improve the baseline performance in NTCD-TIMIT task.

Index Terms— Audio-visual speech recognition, bimodal DFSMN, robust speech recognition, dropout, multi-condition training

1. INTRODUCTION

Although automatic speech recognition (ASR) has achieved great progress in the past few years, the performance of ASR systems in noisy environments still far behind human speech recognition. Traditionally, various front-end signal based or model based speech enhancement techniques [1, 2, 3] are widely used to improve the intelligibility and quality of noisy speech. These techniques significantly improve the performance of back-end acoustic modeling in noisy environments. However, even with these speech enhancement techniques, the performance of ASR systems in noisy environments still can not match the competence and noise robustness of human speech recognition. Clearly, non-traditional approaches, that use orthogonal sources of information to the acoustic input, are needed to achieve ASR performance closer to the human speech perception level, and robust enough to be applied in noisy environments [4].

Both human speech production and perception are bimodal in nature [5]. Especially in noisy environments, humans use visual modality in addition to the traditional audio modality to help deciding what has spoken. Back to 1954, the visual modality benefits to speech intelligibility in noise has been quantified by Sumby and Pollack in [6]. After then, the integration of audio and visual information in perceiving speech has been demonstrated in [7]. Thereby, *lipreading* and *audio-visual speech recognition (AVSR)* [8, 9, 10, 11] that use the visual information in addition to the acoustic input have attracted more and more attention of researchers. Not surprisingly, AVSR systems have been shown to outperform conventional audio-only ASR systems over a wide range of conditions, and the performance gains are particularly impressive in noisy environments.

In AVSR research, the audio-visual integration strategy is one of the most important aspects. Generally speaking, it can be categorized into three major categories: *feature fusion* (FF) (also called early integration) [12, 13, 14, 15, 16], *decision fusion* (DF) (also called late integration) [12] and *intermediate fusion* (IF) [17, 18]. As to feature fusion, audio and visual features are spliced to form a new set of audio-visual features and then used for audio-visual acoustic modeling. It can be simply done by concatenating the raw audio and video feature vectors or after applying some linear or nonlinear transformation. For decision fusion, the audio-only and video-only ASR systems are first trained independently using the audio and visual features, respectively. And then recognition results are combined using approaches like ROVER [19]. Compared to FF, the advantage of DF is that it's able to control the contribution of audio and video modalities to the overall recognition results according to their reliabilities using stream weights. However, the drawback of DF is that it's unable to utilize the temporal correlation between audio and visual modalities. Experimental results in [20] show that DF based AVSR system performs much better in low SNR test sets ($\leq$5dB) while worse in high SNR test sets ($\geq$10dB) when compared to F-F based AVSR system. An additional problem of FF based AVSR systems is that its performance will suffer from huge degradation if the visual information does not exist during decoding. During model training, we have parallel audio and visual training data. But for practical applications, it maybe difficult to capture the video stream of speakers all the time. This mismatch between training and decoding will seriously hurt the performance of an AVSR system.

Another challenge of AVSR research is the lack of publicly available audio-visual large vocabulary continuous speech recognition corpora. Recently, the released TCD-TIMIT [21] and NTCD-TIMIT [22] databases help to alleviate this problem, and the NTCD-TIMIT corpus is adopted in this paper. In this work, we will address the challenges of AVSR through three aspects: 1) optimal integration of acoustic and visual information; 2) robust performance with multi-condition training; 3) robust modeling against missing visual information during decoding. Previous experiments in NTCD-TIMIT [20, 22] have already evaluated the performance of AVSR systems in three setups: a clean-train-clean-test, a clean-train-noisy-test and a matched training setup. As a supplement, we conduct experiments with a multi-condition experimental setup by using an 150-hours multi-condition audio data with various noise

types and SNRs. Moreover, based on the previous FSMN related works [23, 24], we propose a novel *bimodal-DFSMN* to jointly learn audio-visual feature fusion and acoustic modeling, and utilize a per-frame dropout [25] approach to enhance the robustness of AVSR system against the missing of visual modality. Bimodal-DFSMN uses *audio-net* and *visual-net* to perform nonlinear feature transformations for audio and visual streams respectively, and then concatenate these transformed features before fed into a *joint-net*. The use of FSMN-like architecture will help to effectively model the long term dependency as well as the temporal correlation in audio and visual signals. Experimental results show that the proposed bimodal-DFSMN with multi-condition training can significantly improve the performance of baseline systems in the NTCD-TIMIT task. It achieves a phone error rate of 12.9% on clean test set and an average phone error rate of 26.2% on all test sets (clean, different SNRs, different noise types) while the baseline systems in [22] are 20.8% and 52.9%, respectively. Furthermore, we can significantly enhance the robustness of AVSR system to the missing of visual information during decoding by introducing per-frame dropout regularization during model training.

2. BIMODAL DFSMN

Recently, audio-only speech recognition has made a great progress. One of the main reasons is the usage of more powerful neural networks structures, especially those structures that can model long-term dependency in speech signals such as the long short term memory recurrent neural networks (LSTM-RNN) [26], time delay neural network (TDNN) [27], feedforward sequential memory networks (FSMN) [23] and its variant DFSMN [24]. However, so far the commonly-used neural network structure in audio-visual speech recognition (AVSR) is still the simplest deep neural network (DNN). Obviously, the performance of AVSR system can be significantly improved if more powerful acoustic models are used.

In this work, we propose a novel *bimodal-DFSMN* to jointly learn audio-visual feature fusion with acoustic modeling for AVSR. The architecture of bimodal-DFSMN is as shown in Figure 1, which consists of three main components: audio-net, visual-net, and joint-net. The audio-net and visual-net are used to convert acoustic and visual features into a deep representation, respectively. These outputs of audio-net and visual-net are concatenated before feeding into a joint-net. Given input acoustic feature sequence ($\mathbf{x_a}$) and visual feature sequence ($\mathbf{x_v}$), the output of bimodal-DFSMN can be denoted as:

$$\mathbf{y} = f_{joint}(f_{audio}(\mathbf{x_a}); f_{visual}(\mathbf{x_v})). \quad (1)$$

Here, f_{audio}, f_{visual} and f_{joint} denote the transformations of audio-net, visual-net and joint-net, respectively.

In this work, we adopt a same architecture for audio-net and visual-net, which consists of a ReLU layer, a linear layer and four DFSMN components. For the joint net, from bottom to top, it contains a ReLU layer, a linear layer, four DFSMN components and two ReLU layers. The ReLU layer denotes a linear transformation followed by a ReLU activation function. Detailed composition of DFSMN component is as shown in Figure 1, which consists of four parts: a ReLU layer, a linear layer, a memory block and a skip connection from the bottom memory block, except for the first one that without the skip connection from the bottom layer. The operation of the ℓ-th DFSMN component take the following form:

$$\mathbf{h}_t^\ell = \max(\mathbf{W}^\ell \mathbf{m}_t^{\ell-1} + \mathbf{b}_t^\ell, 0) \quad (2)$$

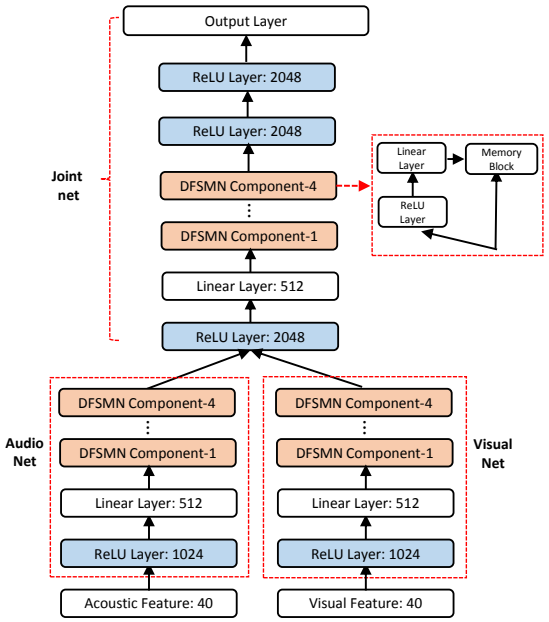

Fig. 1. Illustration of bimodal DFSMN.

$$\mathbf{p}_t^\ell = \mathbf{V}_t^\ell \mathbf{h}_t^\ell + \mathbf{v}_t^\ell \quad (3)$$

$$\mathbf{m}_t^\ell = \mathbf{m}_t^{\ell-1} + \mathbf{p}_t^\ell + \sum_{i=0}^{N_1^\ell} \mathbf{a}_i^\ell \odot \mathbf{p}_{t-s_1*i}^\ell + \sum_{j=1}^{N_2^\ell} \mathbf{c}_j^\ell \odot \mathbf{p}_{t+s_2*j}^\ell \quad (4)$$

Here, $\mathbf{h}_t^\ell$ and $\mathbf{p}_t^\ell$ denote the outputs of the ReLU layer and linear layer respectively. $\mathbf{m}_t^\ell$ denotes the output of the ℓ-th memory block. N_1^ℓ and N_2^ℓ denotes the look-back and lookahead orders of the ℓ-th memory block, respectively. s_1 is the stride factor of look-back filter and s_2 is the stride of lookahead filter. The detailed structure of DFSMN components in audio-net and visual-net is: ReLU layer with 1024 units, linear layer with 512 units, memory block with $N_1^\ell = 5$, $N_2^\ell = 5$, $s_1 = 1$ and $s_2 = 1$. And the detailed structure of DFSMN component in joint-net is: ReLU layer with 2048 units, linear layer with 512 units, memory block with $N_1^\ell = 20$, $N_2^\ell = 20$, $s_1 = 2$ and $s_2 = 2$. The use of memory blocks will enable the bimodal-DFSMN to effectively model the long-term dependency in both audio and video signals.

3. PER-FRAME DROPOUT FOR VISUAL MODALITY MISSING PROBLEM

Dropout is a powerful technology introduced in [28] for improving generalization capability of neural networks. During training, dropout can reduce overfitting by randomly omitting a fraction of units in all layers on each training case to prevent co-adaptation. Previous works shown that dropout plays a big role in computer vision tasks that use CNN-type networks. As to speech recognition with LSTM-type networks, the original per-element dropout doesn't work well. As an alternative, the per-frame dropout is proposed in [25] that performs well for various LVCSR tasks with TDNN-LSTM networks. Unlike per-element dropout in which each element of the dropout mask is chosen independently, in per-frame dropout the dropout mask vector is set to either zero or one. For a time in-

Table 1. Detailed experimental results for various models on NTCD-TIMIT corpus.

Exp	Model	Training Data	Audio	Video	PER(%)						
					Clean	20dB	15dB	10dB	5dB	0dB	**AVG**
1	DNN	Clean-5-hours	ON	OFF	21.6	46.6	57.0	67.6	76.8	85.3	**59.1**
2	DNN	Clean-5-hours	OFF	ON	65.3						
3	DNN-FF	Clean-5-hours	ON	ON	20.8	39.1	48.7	59.2	70.1	79.6	**52.9**
4	DNN-DF	Clean-5-hours	ON	ON	22.4	47.0	56.4	62.1	64.2	65.5	**52.9**
5	DFSMN	Clean-5-hours	ON	OFF	18.7	41.2	51.3	61.7	70.7	77.3	**53.5**
6	DFSMN	Clean-5-hours	OFF	ON	62.5						
7	DFSMN-FF	Clean-5-hours	ON	ON	17.4	34.4	42.8	52.8	62.3	70.5	**46.7**
8	DFSMN-DF	Clean-5-hours	ON	ON	19.4	41.7	51.5	58.5	60.3	62.6	**49.0**
9	Bimodal-DFSMN	Clean-5-hours	ON	ON	16.3	34.7	43.9	54.4	63.8	71.4	**47.4**
10	DFSMN	MC-150-hours	ON	OFF	14.5	20.8	26.9	36.6	50.8	64.9	**35.8**
11	DFSMN-FF	MC-150-hours	ON	ON	13.7	19.0	22.9	29.0	37.9	48.8	**28.6**
12	DFSMN-DF	MC-150-hours	ON	ON	15.5	21.4	27.4	37.0	51.2	57.9	**35.1**
13	Bimodal-DFSMN	MC-150-hours	ON	ON	12.9	17.7	21.3	26.7	34.4	44.0	**26.2**

stance $\mathbf{x}_t$, the per-frame dropout can be expressed as:

$$f_{dropout,p}(\mathbf{x}_t) = \begin{cases} \mathbf{x}_t & \alpha \geq p \\ \mathbf{0} & \alpha < p \end{cases}. \quad (5)$$

Here, α is a Bernoulli random scalar and p is the dropout probability.

For feature fusion (FF) based AVSR, during model training, we use the parallel audio and video corpus. However, during model testing in practical applications, the visual modality missing problem may occur due to the difficulty to capture speaker's mouth area all the time. This mismatch problem between training and test will cause great damage to the performance. In this work, we use the per-frame dropout regularization to improve the robustness of AVSR system and and to deal with this problem. We adopt a per-frame dropout operation for the visual-net in bimodal-DFSMN. As to imitate the missing of visual modality, we use the per-frame dropout for the input layer of visual-net. Given input acoustic feature sequence ($\mathbf{x_a}$) and visual feature sequence ($\mathbf{x_v}$), the operation in bimodal-DFSMN with per-frame dropout can be denoted as:

$$\mathbf{y} = f_{joint}(f_{audio}(\mathbf{x_a}); f_{visual}(f_{dropout,p}(\mathbf{x_v}))). \quad (6)$$

In experiments, we will investigate the performance of the bimodal-DFSMN AVSR system with various missing ratios of visual modality. For example, if the per-frame dropout probability p is set to be 0.5, it means 50% of the input visual features are randomly set to zero.

4. EXPERIMENTS

In this section, we investigate the performance of bimodal-DFSMN with multi-condition training and per-frame dropout regularization for audio-visual speech recognition on NTCD-TIMIT corpus.

4.1. Experimental Setups

NTCD-TIMIT [22] is a newly published audio-visual speech recognition corpus based on the TCD-TIMIT [21] corpus. It contains the audio signals and the visual features of 56 Irish speakers from the TCD-TIMIT database. In addition to a down-sampled version of the clean TCD-TIMIT utterances, 36 noisy versions ({Six noise types: white, babble, car, living room, cafe, street} $\times$ { six types of SNR: 20dB, 15dB, 10dB, 5dB, 0dB, -5dB }) have been created. Both the clean and noisy signals of NTCD-TIMIT are sampled at a sampling rate of 16 kHz. Following TCD-TIMIT corpus, the clean set of NTCD-TIMIT is divided into training set, development set and test set. The training set consists of 39 speakers with about 5 hours data. And the development set contains 8 speakers with about 1 hour data for hyper parameters tuning. Evaluation is performed in term of phone error rate (PER) on the 9-speaker test set (about 1.2 hours of data). The NTCD-TIMIT database also contains Kaldi scripts for training and decoding audio-only, video-only, and audio-visual ASR models. Experimental results obtained using these scripts are detailed in [22], which are used as the baseline systems in this study.

We have evaluated the proposed approaches on NTCD-TIMIT corpus with two experimental setups: 1) Clean-5-hours setup; 2) MC-150-hours multi-condition setup. The *Clean-5-hours* setup is the same to the original clean setup in [22] that uses the 5 hours clean training data. As to the *MC-150-hour* setup, we mix the training data of 30 noisy versions that consists of six noise types with five level SNR (20dB, 15dB, 10dB, 5dB, 0dB), resulting in a multi-condition training data. Models are evaluated on the clean test set and 30 noisy test sets. For these noisy test sets we report the average PER of six noise types for each SNR.

We follow the front-end processing in [22] for acoustic and visual feature extraction. The visual front-end contains three steps: face and ROI detection, ROI post-processing, visual feature extraction. The difference between the acoustic frame rate (100 frame/s) and the visual frame rate (30 frame/s) is compensated by repeating visual frames according to the digital differential analyzer algorithm. As a result, both the acoustic and visual features are the 40-dimensional fMLLR feature at a frame rate of 100 frame/s.

4.2. Experimental Results

4.2.1. Baseline systems

We have reproduced the audio-only, video-only and audio-visual baseline systems using the released Kaldi scripts in [22]. The speaker-independent acoustic DNN-HMM hybrid models have been trained using the frame-state alignments obtained by applying the forced alignment algorithm to the SAT tri-phone GMM-HMM models. The DNN has 6 hidden layers, each of which consists of 1024 neurons with sigmoid activation functions. The number of units in the output softmax layer is 1975, which is the number of the tied tri-phone states. The DNN is first trained using the frame-level

cross-entropy (CE) objective function and further optimized with state-level minimum Bayes risk (sMBR) criterion. The audio-only and video-only systems are trained using the 40-dimensional fMLLR features extracted from the original 5 hours of clean audio data and video data, respectively. For audio-visual baseline systems, we have trained DNNs with feature fusion (DNN-FF) and decision fusion (DNN-DF). The detailed experimental results of these baseline systems are as shown in Table 1 (Exp1-4).

4.2.2. Acoustic model architecture and fusion method

In this experiment, we investigate the acoustic model architecture and fusion method to the performance of ASR and AVSR systems. Firstly, we replace the baseline DNN with DFSMN and evaluate the performance of audio-only, video-only, and audio-visual systems with more powerful acoustic model using the *Clean-5-hours* setup. The architecture of DFSMN is the same to the *joint-net* in bimodal DFSMN as introduced in Sec.2. Experimental results in Table 1 (Exp 5-8) show that DFSMN based systems (audio-only, video-only and audio-visual) can significantly outperform the corresponding DNN baseline systems. For example, the DFSMN-FF AVSR system achieves an average PER of 46.7% that obtains 11.7% relative PER reduction compared to the DNN-FF AVSR system.

Comparison of DFSMN based AVSR systems with feature fusion (FF) and decision fusion (DF) shows that DFSMN-FF performs much better in low SNR ($\leq$ 5dB) test sets while worse in high SNR ($\geq$ 10dB) test sets than DFSMN-DF, which is consistent to the experimental phenomena of baseline DNN-FF and DNN-DF systems. Unlike FF, DF can control the contribution of audio and video streams to the overall recognition results according to their reliabilities using stream weights. As a result, the performance of DF based AVSR system is close to the video-only system in low SNR test set since it is focused on the video signals in low SNR. On the other hand, DFSMN-DF performs worse than DFSMN-FF in high SNR test sets since DF can not utilize the temporal correlation between the audio and visual modalities.

We have also evaluated the performance of bimodal-DFSMN based AVSR system (Exp 9 in Table 1) using the Clean-5-hours experimental setup. Compared to DFSMN-FF (Exp 7), bimodal-DFSMN performs much better in clean test set while worse in noisy test sets. This may due to bimodal-DFSMN is a more powerful model that is easily overfitting to the training data. As a conjecture, if we can effectively extend the training data, then bimodal-DFSMN will show its model capacity. Accordingly, we construct the MC-150-hours multi-condition experimental setup to evaluate this conjecture.

4.2.3. Multi-condition training

In this experiment, we have trained DFSMN, DFSMN-FF, DFSMN-DF and bimodal-DFSMN with the MC-150-hours experimental setup. Detailed experimental results are listed in Table 1 (Exp 10-13). Compared to the performance of systems trained with the Clean-5-hours experimental setup, all types of systems can achieve a significant improvement not only in the clean test set but also in the noisy test set. On average, the relative performance improvements of DFSMN, DFSMN-FF, DFSMN-DF, and bimodal-DFSMN are 33.1%, 38.8%, 28.4% and 44.7%, respectively. In line with our previous conjecture, the bimodal-DFSMN shows its capacity by increasing the coverage of training data. Comparison of Exp 13 and Exp 11 demonstrates that bimodal-DFSMN can outperform the DFSMN-FF in all test sets. From Exp 10-13, we can also see that even the audio-only ASR system performs better than the video-only ASR system in test sets with different SNR, the audio-visual fusion with FF or bimodal-DFSMN can still significantly improve the performance of AVSR systems. However, the DF based system doesn't work well since the performance of video-only ASR system is poor.

Table 2. Performance (PER in %) of bimodal-DFSMN AVSR systems trained with various per-frame dropout probabilities (Train-p) and tested with various missing ratios of visual modality (Test-p).

Train-p	Test-p	Clean	20dB	15dB	10dB	5dB	0dB	AVG
0.0	**0.0**	**12.9**	**17.7**	**21.3**	**26.7**	**34.4**	**44.0**	**26.2**
	0.3	14.2	19.8	24.0	30.6	39.8	50.3	29.8
	0.5	16.8	24.4	30.1	38.5	48.8	59.1	36.3
	0.8	29.3	46.6	55.3	64.5	72.1	77.5	57.6
	1.0	48.2	67.3	72.8	77.6	81.0	83.6	71.8
0.3	0.0	13.6	18.6	22.5	28.8	37.9	48.8	28.4
	0.3	**13.2**	**18.1**	**22.1**	**28.2**	**37.5**	**49.5**	**27.9**
	0.5	13.3	18.2	22.2	28.6	38.0	49.4	28.3
	0.8	16.2	22.5	27.6	35.6	46.6	58.2	34.5
	1.0	31.0	45.1	53.0	61.7	50.9	78.9	56.8
0.5	0.0	13.3	18.8	22.8	29.1	38.5	49.6	28.7
	0.3	12.9	18.2	22.0	28.0	37.6	48.6	28.0
	0.5	**12.8**	**17.9**	**21.8**	**28.0**	**37.5**	**48.9**	**27.8**
	0.8	12.9	19.7	24.2	31.4	41.9	53.9	30.7
	1.0	27.1	41.7	50.0	59.5	68.8	76.2	53.9
0.8	0.0	14.3	20.1	24.7	31.8	41.7	52.9	30.9
	0.3	13.3	18.5	22.4	28.7	38.4	49.8	28.5
	0.5	12.8	17.6	21.3	28.8	38.4	49.8	28.1
	0.8	**12.6**	**17.1**	**20.8**	**27.0**	**36.6**	**48.9**	**27.2**
	1.0	17.3	27.4	35.3	46.4	59.8	71.2	42.9

4.2.4. Per-frame dropout for visual modality missing problem

Experimental results in Sec.4.2.3 indicate that feature fusion seems to be a better choice than decision fusion when the AVSR models are trained using the multi-condition experimental setup. For feature fusion, one challenge is how to use the FF-based AVSR systems when the visual information does not exist during decoding. As introduced in Sec.3, we propose to handle this problem by using the *per-frame dropout*. We have trained bimodal-DFSMN with various per-frame dropout probabilities (Train-p) and tested these models with various missing ratios of visual modality (Test-p). Detailed experimental results are as shown in Table 2. For the baseline system without using per-frame dropout (Train-p =0), performance will drop significantly as the increasing of missing ratios of visual modality during testing. Results in Table 2 also show that models trained with per-frame dropout (Train-p: 0.3, 0.5, 0.8) are more robust to the missing of visual modality. For example, in a mismatch test, model with Train-p =0.8 and Test-p =0.3 can still achieve a promising result. The performance of all models with Test-p =1.0 is very poor, which indicates the importance of visual modality in AVSR systems.

5. CONCLUSIONS

In this work, we have conducted audio-visual speech recognition by addressing the challenges through three aspects: 1) optimal integration of acoustic and visual information; 2) robust performance with multi-condition training; 3) robust modeling against missing visual information during decoding. Experimental results show that the proposed bimodal-DFSMN with multi-condition training can significantly improve the performance of AVSR system. Furthermore, the per-frame dropout can enhance the robustness of bimodal-DFSMN to the missing of visual modality during testing. In the experiments, the performance of video-only ASR system is not very good. Further improvement of AVSR performance can be obtained through more powerful visual front-end processing and modeling methods.

6. REFERENCES

[1] Pascal Chevalier and Audrey Blin, "Widely linear MVDR beamformers for the reception of an unknown signal corrupted by noncircular interferences," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5323–5336, 2007.

[2] DeLiang Wang, Ulrik Kjems, Michael S Pedersen, Jesper B Boldt, and Thomas Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336–2347, 2009.

[3] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[4] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, pp. 23, 2004.

[5] David G Stork and Marcus E Hennecke, *Speechreading by humans and machines: models, systems, and applications*, vol. 150, Springer Science & Business Media, 2013.

[6] William H Sumby and Irwin Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[7] Harry McGurk and John MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746, 1976.

[8] Aggelos K Katsaggelos, Sara Bahaadini, and Rafael Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, 2015.

[9] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[10] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas, "Lipnet: Sentence-level lipreading," *arXiv preprint*, 2016.

[11] Hendrik Meutzner, Ning Ma, Robert Nickel, Christopher Schymura, and Dorothea Kolossa, "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5320–5324.

[12] Pascal Teissier, Jordi Robert-Ribes, J-L Schwartz, and Anne Guérin-Dugué, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 629–642, 1999.

[13] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[14] Jing Huang and Brian Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7596–7599.

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

[16] Di Hu, Xuelong Li, et al., "Temporal multimodal learning in audiovisual speech recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3574–3582.

[17] Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea Kolossa, "Learning dynamic stream weights for coupled-hmm-based audio-visual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 863–876, 2015.

[18] Simon Receveur, Robin Weiß, and Tim Fingscheidt, "Turbo automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 5, pp. 846–862, 2016.

[19] Jonathan G Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[20] Ahmed Hussen Abdelaziz, "Comparing fusion models for DNN-based audiovisual continuous speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 3, pp. 475–484, 2018.

[21] Naomi Harte and Eoin Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[22] Ahmed Hussen Abdelaziz, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," in *Proc. Interspeech*, 2017, pp. 3752–3756.

[23] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu, "Nonrecurrent neural structure for long-term dependence," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 871–884, 2017.

[24] Shiliang Zhang, Ming Lei, Zhijie Yan, and Lirong Dai, "Deep-FSMN for large vocabulary continuous speech recognition," *arXiv preprint arXiv:1803.05030*, 2018.

[25] Gaofeng Cheng, Vijayaditya Peddinti, Daniel Povey, Vimal Manohar, Sanjeev Khudanpur, and Yonghong Yan, "An exploration of dropout with LSTMs," in *Proc. Interspeech*, 2017.

[26] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[27] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*, 2015.

[28] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.