2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Robust Audio-Visual Speech Recognition Based on Hybrid Fusion

Hong Liu, Wenhao Li, Bing Yang
Key Laboratory of Machine Perception
Shenzhen Graduate School, Peking University
Shenzhen, China
{hongliu, wenhaoli}@pku.edu.cn, bingyang@sz.pku.edu.cn

***Abstract*—The fusion of audio and visual modalities is an important stage of audio-visual speech recognition (AVSR), which is generally approached through feature fusion or decision fusion. Feature fusion can exploit the covariations between features from different modalities effectively, whereas decision fusion shows the robustness of capturing an optimal combination of multi-modality. In this work, to take full advantage of the complementarity of the two fusion strategies and address the challenge of inherent ambiguity in noisy environments, we propose a novel hybrid fusion based AVSR method with residual networks and Bidirectional Gated Recurrent Unit (BGRU), which is able to distinguish homophones in both clean and noisy conditions. Specifically, a simple yet effective audio-visual encoder is used to map audio and visual features into a shared latent space to capture more discriminative multi-modal feature and find the internal correlation between spatial-temporal information for different modalities. Furthermore, a decision fusion module is designed to get final predictions in order to robustly utilize the reliability measures of audio-visual information. Finally, we introduce a combined loss, which shows its noise-robustness in learning the joint representation across various modalities. Experimental results on the largest publicly available dataset (LRW) demonstrate the robustness of the proposed method under various noisy conditions.**

***Index Terms*—Audio-Visual Fusion, Robust Speech Recognition, Multi-modality, Hybrid Fusion**

I. Introduction

Automatic speech recognition (ASR) has been applied to a wide range of human-robot interaction systems such as service robots, mobile phones, etc. Since audio-based speech recognition can be easily affected by acoustic noise, audio-visual speech recognition (AVSR) introduces visual speech information to improve the robustness of speech recognition, which has aroused wide research attention in the past decades [1]–[3]. Despite the encouraging results of AVSR, it remains a challenging problem to fuse the two modalities more effectively and robustly due to the intrinsically ambiguous nature of homophones, especially in noisy environments [4].

Traditional AVSR methods first extract the audio and visual features and then put them into a classifier with feature fusion or decision fusion [5]–[7]. A parallel two-step keyword spotting strategy based on decision fusion was proposed to combine the audio information and visual information to enhance the audio-visual keyword spotting system based on Hidden Markov Model [8]. Recently, deep learning approaches have been widely used to solve the AVSR problem [9], [10]. The recurrent temporal multi-modal restricted Boltzmann machine (RTMRBM) was introduced to model multi-modal sequences in the task of AVSR [11]. Several end-to-end works have been presented and extended to audio-visual models. Chung *et al.* used an end-to-end model with an attention mechanism to recognize phrases and sentences [12]. Petridis *et al.* proposed an end-to-end audio-visual model with feature fusion which learns to extract features directly from both the pixels and spectrograms [13].

Although the promising results have been made by recent works [14]–[16], we observe that most of the existing works only focus on either feature fusion or decision fusion, while ignoring the complementarity between these two fusion strategies [17]. Specifically, feature fusion can effectively exploit the covariations between features from different modalities to learn more discriminative representation, while utilizing decision fusion shows the robustness of capturing an optimal combination of two modalities, especially in different noisy conditions [18]. These observations motivate us to develop a method that can effectively embed both fusion strategies into a hybrid fusion architecture, and leverage it for AVSR to resolve the challenging issue of inherent ambiguity in noisy environments.

In this paper, a novel audio-visual speech recognition method based on hybrid fusion is proposed. Similar to the work of [19], features of each modality are extracted from the input audio waveforms and mouth regions using ResNet-based audio and visual encoders, followed by a 2-layer Bidirectional Gated Recurrent Unit (BGRU) to model the temporal dynamics, respectively. Then an audio-visual encoder is applied to perform feature fusion for both streams and fed to another 2-layer BGRU. Finally, the losses of audio, visual, and audio-visual streams are combined to train the network end-to-end, and the final predictions are made through a decision fusion module. Note that different from the end-to-end model proposed in [19], our work has three improvements as follows:

- A simple yet effective audio-visual encoder is used to fuse the audio and visual features into a discriminative multi-modal feature rather than concatenate them directly;
- A combined loss is introduced to learn the joint representation across audio-visual modalities robustly instead of using one audio-visual loss;
- A hybrid fusion architecture is adopted to combine the

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412817

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:33:48 UTC from IEEE Xplore. Restrictions apply.

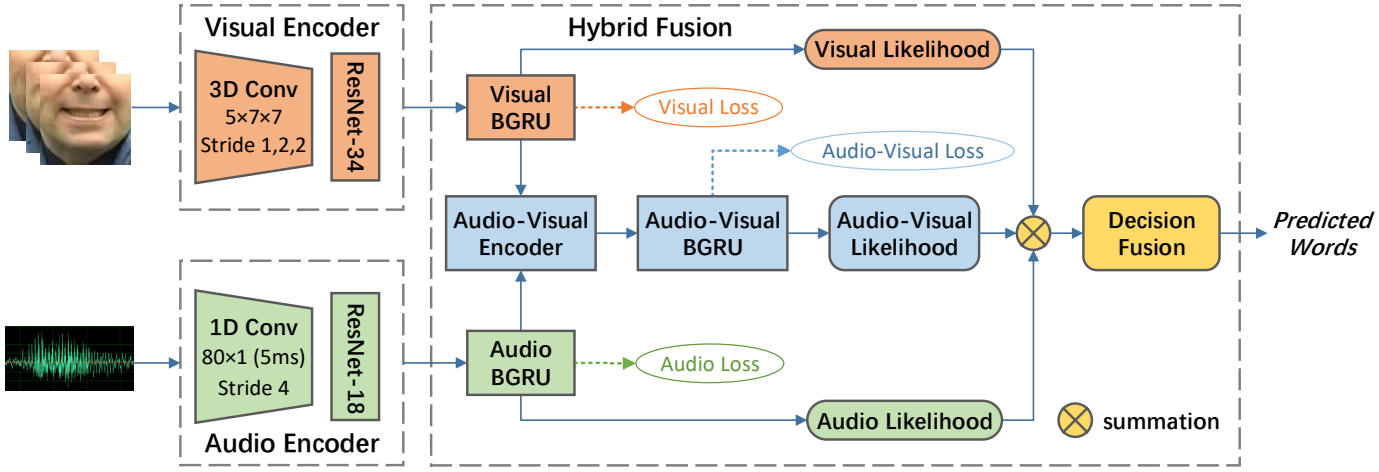


Fig. 1: Overview of the proposed end-to-end audio-visual speech recognition method based on hybrid fusion.

information of both modalities effectively and robustly, which is capable of adapting to various noisy conditions.

Experimental results on the LRW dataset [20] demonstrate that the proposed method outperforms the state-of-the-art methods in different noisy environments with a margin of 2.88% under noise level of -5 dB and 6.88% at -10 dB of babble noise.

II. Methodology

The overview of the proposed audio-visual model is shown in Fig. 1, which is composed of three components: audio network, visual network, and audio-visual hybrid fusion network. Audio waveforms and mouth region of interest (ROI) sequences are put into the audio and visual networks to extract features respectively and then integrated with an audio-visual encoder layer. A hybrid fusion architecture is introduced to combine audio and visual information in both feature and decision levels for the audio-visual fusion network.

A. Audio Network

The audio network contains an audio encoder and an audio BGRU as shown in the lower part of Fig. 1. The audio encoder includes an 18-layer ResNet with 1D kernels as the audio waveforms are 1D signals, followed by two BGRU layers (audio BGRU). The audio waveforms are put into the first convolutional layer with a temporal kernel of 5ms and a stride of 0.25ms to reduce the dimensionality of time-frequency and then fed to the following ResNet with the kernel size of 3×1. Next, the output of the ResNet with 29 frames is fed to a 2-layer BGRU with 1024 cells to model the long-term and short-term dependency patterns and learn the temporal information for audio modality. Finally, a softmax output layer is used to get the audio likelihood $P\left(c_i^a|x^a\right)$, where x represents the given sequences and c_i is the i-th isolated words.

B. Visual Network

Following the similar architecture with the audio network, the visual network is composed of a visual encoder and a visual BGRU as shown in the upper part of Fig. 1. The visual encoder includes a spatial-temporal convolutional layer followed by a 34-layer ResNet. The spatial-temporal convolutional layer is capable of capturing the short-term dynamic nature of the mouth regions and is confirmed to perform better than aggregating spatial-only features [21]. 3D CNN is utilized to learn the spatial and temporal features of the mouth ROI sequences through the 3D convolutional operation. Firstly, the mouth ROI sequences are put into a spatial-temporal convolutional layer with 64 3D kernels of $5 \times 7 \times 7$ size to capture the short-term dependency patterns and then fed to a 34-layer ResNet. Furthermore, the visual BGRU including two independent BGRU layers with 1024 cells is employed to model the temporal dynamics. Finally, a softmax output layer is used to calculate the visual likelihood $P\left(c_i^v|x^v\right)$.

C. Audio-Visual Hybrid Fusion Network

The complementarity of audio and video information can be used to improve the performance of the AVSR system and audio-visual fusion is generally approached through feature fusion or decision fusion. In this work, instead of investigating which approach is better, our key insight is to make the two paradigms form a strong collaboration. Note that different from the recent paper [22] that uses a joint CTC/attention hybrid architecture for AVSR, our work considers a hybrid fusion architecture that combines feature fusion and decision fusion to improve the effectiveness and robustness of the AVSR system, as depicted in Fig. 1.

1) Audio-Visual Feature Fusion: Different from existing works [19], [23] which directly concatenate the features of the two modalities into a high-dimensional vector, we introduce a simple yet effective audio-visual encoder to map audio and visual features into a shared latent space to capture more discriminative multi-modal feature. This light-weight encoder, which contains only two fully-connected layers with batch normalization, dropout, and Rectified Linear Units (ReLU) activation function, as well as residual connection, is able to learn more crucial information and find the internal correlation

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:33:48 UTC from IEEE Xplore. Restrictions apply.

Fig. 2: **Top:** Example of video frames from Lip Reading in the Wild dataset. **Bottom:** Mouth ROI sequences for 'about' from two different speakers.

between spatial-temporal information from audio and visual networks effectively.

The output features from audio and visual BGRU are fed to an audio-visual encoder in order to fuse the information from the two modalities and then put into a 2-layer BGRU which consists of 1024 cells in each layer (using the same architecture of [19]).

2) Audio-Visual Decision Fusion: Decision fusion is one of the most promising solutions for audio-visual models in handling different noisy environments [24]. In this work, decision fusion is employed for inference in order to complementarily utilize the reliability measures of audio and visual information and adapt to diverse noisy environments.

To calculate the audio-visual likelihood, a fully-connected layer is first used to compress the output units to the number of isolated words. Then, a softmax layer is adopted to get the likelihood of audio-visual network $P\left(c_i^{av}|x^{av}\right)$. Finally, the fusion likelihood from the audio-visual decision fusion module is calculated as the summation of likelihoods from the audio, visual and audio-visual network:

$$P\left(c_i|x\right) = \alpha P\left(c_i^a|x^a\right) + \beta P\left(c_i^v|x^v\right) + \gamma P\left(c_i^{av}|x^{av}\right), \quad (1)$$

where $P\left(c_i^a|x^a\right)$, $P\left(c_i^v|x^v\right)$ and $P\left(c_i^{av}|x^{av}\right)$ denotes the likelihood from audio, visual and audio-visual networks (without decision fusion module), respectively. α, β and γ are hyper-parameters to control the weighting factor for the audio, visual and audio-visual information, and satisfy the constraints:

$$\alpha + \beta + \gamma = 1, 0 \leq \alpha, \beta, \gamma \leq 1. \quad (2)$$

The class label inferred by audio-visual fusion network is obtained through:

$$z = \arg\max_i \left\{P\left(c_i|x\right)\right\}. \quad (3)$$

Similarly, the class label of audio-only, visual-only, and audio-visual networks (without decision fusion module) can be estimated by the maximizing operation for $P\left(c_i^a|x^a\right)$, $P\left(c_i^v|x^v\right)$ and $P\left(c_i^{av}|x^{av}\right)$, respectively.

3) Audio-Visual Training: In order to improve the complementation between audio and visual information and learn the joint representation robustly for the audio-visual model, the losses of three modalities are combined and each network produces a cross-entropy loss component as:

$$L_k = -\sum_{i=1}^{C} y_i \log P\left(c_i^k|x^k\right), \quad (4)$$

where $k \in \{a, v, av\}$ denotes audio, visual and audio-visual stream, respectively, y indicates the true class label of each sequence and C is the number of target isolated words.

In our implementation, we first train the audio network, visual network, and audio-visual network separately with audio loss L_a, visual loss L_v, and audio-visual loss L_{av}, respectively. Then, the entire network is fine-tuned in an end-to-end manner to optimize the total objective, which is a weighted sum of the following losses:

$$L = \lambda_a L_a + \lambda_v L_v + \lambda_{av} L_{av}, \quad (5)$$

where λ is the weighting factor, L denotes the combined loss of the audio-visual network. This loss function ensures that the model focuses on various modalities and shows its noise-robustness in learning the joint representation across audio-visual modalities.

III. EXPERIMENTS AND ANALYSES

A. Dataset

The dataset used in our experiments is the Lip Reading in the Wild (LRW) dataset [20], which is the largest public dataset for AVSR in English. The LRW dataset consists of short video clips with 29 frames (1.16 seconds) from BBC television as shown in Fig. 2, with more than 1000 speakers saying 500 different isolated words, such as "YOUNG", "SOCIAL" and "UNITED". For each word, there are 800 to 1000 sequences in the training set and 50 sequences in the validation and test sets, respectively. In this work, we follow the same training and evaluation protocols used in

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:33:48 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Word accuracy (%) of the audio-only, visual-only and audio-visual models on the LRW dataset in clean audio condition.

Modalities	Methods	Word accuracy (%)
Audio	Petridis *et al.* [19]	97.70
	Stafylakisa *et al.* [25]	97.96
Visual	Chung *et al.* [20]	61.10
	Chung *et al.* [12]	76.20
	Petridis *et al.* [19]	82.00
	Stafylakis *et al.* [23]	83.00
	Wang *et al.* [26]	83.34
	Zhao *et al.* [27]	84.41
Audio-Visual	Petridis *et al.* [19]	98.20
	Ours	**98.91**

[19], [23], training on 488766 examples, validating on 25000 examples, and testing on 25000 examples. Since the mouth ROIs are already centered, a fixed bounding box of 96×96 is performed to extract them. The audio waveforms are extracted from the videos at the rate of 16kHz for all examples.

B. Experimental Setting

In this work, our implementation is based on the PyTorch framework using one GeForce RTX 2080 Ti GPU with CUDA 10.1. The network is trained using Adam optimizer with a batch size of 32 for all the experiments. An initial learning rate of 0.0003 is used and decreased by 68.3% every time when the training loss does not decrease for every 50 iterations. For the hyper-parameters, we set the α, β and γ as 0.3, 0.3 and 0.4, respectively, $\lambda_a = 0.5$, $\lambda_v = 0.5$ and $\lambda_{av} = 1$ by optimizing on the validation set in our implementation.

Two different noise types from the Noisex92 dataset [28], namely babble and white noises, are adopted to add to the original audio waveforms with different signal-to-noise ratios (SNRs) in order to investigate the robustness of our proposed audio-visual fusion method in noisy environments. To improve the adaptability of our model to audio noise, different SNR levels of noise (between -5 dB to 20 dB, with an interval of 5 dB) is added to the original audio waveforms in training, while the robustness of our model in noisy conditions is evaluated by applying additive noises with various SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5dB and -10 dB to the validation and test sets.

C. Comparison with the State-of-the-art

Since many previous works only focus on lip reading task (visual-only) on the LRW dataset as well as there are few previous audio-only and audio-visual results, we first compare the performance with these works as shown in Table I and our audio-visual model improves the word recognition accuracy to 98.91% on the LRW dataset in clean audio condition.

Next, the performance of our audio-visual model is compared with the method in [19] to evaluate the effectiveness and robustness of our proposed method. Except for babble noise used in [19], white noise is also added to the original audio waveforms. Empirically, we observed that the performance of the audio-only model is very low (only a few percent) when the SNR $\leq$ -10 dB, the audio-only model can't provide useful information to the audio-visual model, so the noises are added to the original audio waveforms with different SNR levels of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB and -10 dB for the validation and test sets. For a fair comparison, the baseline model is fine-tuned on the same training data which is used in our experiments.

Results of the audio-only, visual-only, and audio-visual models are shown in Fig. 3. The accuracy of the visual-only model remains constant since visual information is not affected by the addition of audio noise, while the accuracy of the audio-only model decreases with the existence of stronger noises. It can be seen that our approach outperforms the state-of-the-art method under different levels of babble noise and white noise from -10 dB to 20 dB.

It can be observed that the audio-only model is most affected by babble noise under high-level SNRs, which gets worse performance than white noise. The word recognition accuracy of the audio-visual model is improved from 98.20% to 98.81% when SNR = 20 dB. This improvement in low levels of noise condition is limited (0.61%) as expected since the baseline model is already capable to capture crucial information under low-level noise. However, compared to the state-of-the-art method, it should be noted that our method significantly improves the performance under strong noisy conditions (SNR $\leq$ 5 dB), results in an absolute improvement of 2.88% at -5 dB and 6.88% at -10 dB of babble noise. Note that the baseline model gets worse performance than the visual-only model at -10 dB, however, our audio-visual model outperforms the visual-only model under different levels of noises which indicates that our proposed model is more robust under strong noisy conditions. This is owing to the effective and robust combination of audio and visual information through the proposed hybrid fusion based audio-visual model.

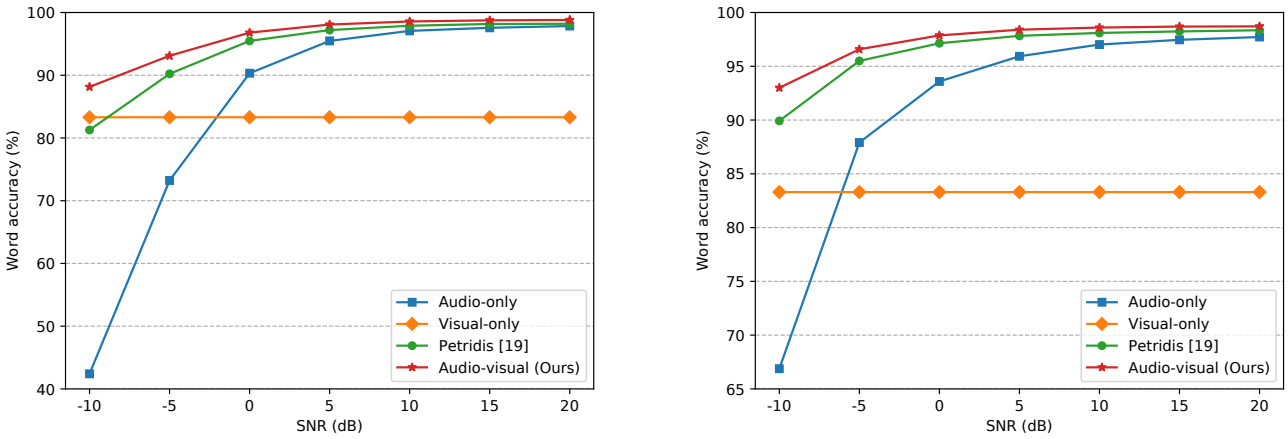
D. Ablation Studies

1) Effect of different components: We study the effect of introducing the audio-visual encoder, the combined loss, and the decision fusion module to our proposed audio-visual model. [19] is used as a baseline which is the state-of-the-art method on LRW. The baseline with four additional processings are taken for comparison.

- w/ AV encoder: insert an audio-visual encoder into baseline.
- w/ combined loss: insert a combined loss into baseline.
- w/ AV decision: insert a decision fusion module into baseline.
- w/ A+V decision: insert a decision fusion strategy which used in [24] into our audio and visual networks.

The results are presented in Table II. We report the performance of different components of our audio-visual models at varying SNR levels of babble noise. It can be seen that the audio-visual encoder (w/ AV encoder) improves the word

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:33:48 UTC from IEEE Xplore. Restrictions apply.



(a) Babble noise

(b) White noise

Fig. 3: Comparisons of the word accuracy (%) with the state-of-the-art method under different levels of babble noise and white noise on LRW dataset.

TABLE II: Ablation studies on different components of our audio-visual model at varying SNR levels of babble noise. Word accuracy (%) on the LRW dataset.

SNR(dB)	-10	-5	0	5	10	15	20
Baseline [19]	81.27	90.22	95.47	97.20	97.89	98.17	98.20
w/ combined loss	86.08	92.58	96.29	97.50	97.97	98.23	98.29
w/ AV encoder	85.05	91.90	96.34	97.81	98.35	98.52	98.62
w/ AV decision	85.91	92.09	96.32	97.80	98.34	98.55	98.64
w/ A+V decision [24]	84.69	89.48	94.64	96.72	97.59	97.98	98.15
Ours	**88.15**	**93.10**	**96.79**	**98.08**	**98.57**	**98.75**	**98.81**

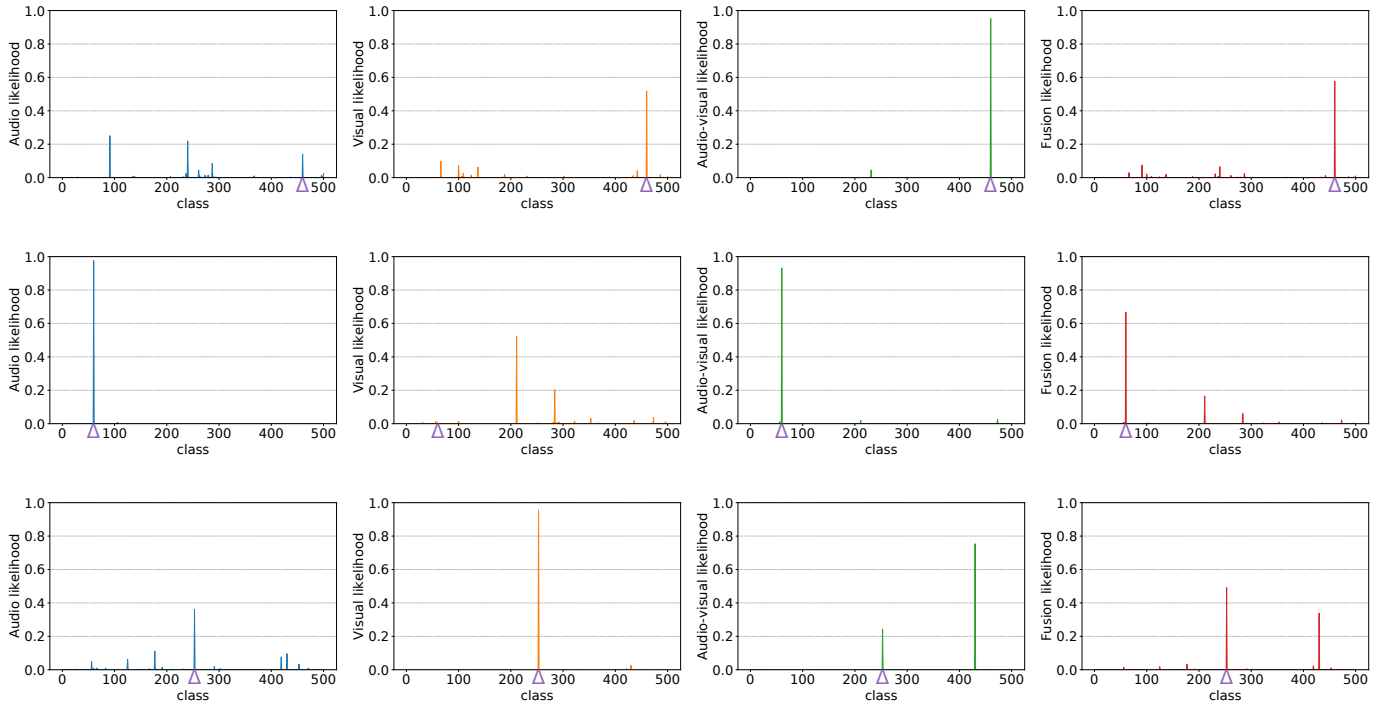
TABLE III: Word accuracy (%) for different hyper-parameter values of audio-visual decision module at varying SNR levels of babble noise on the LRW dataset. The best score for each SNR condition is marked in bold.

α	β	γ	-10	-5	0	5	10	15	20	Average
0.2	0.2	0.6	86.76	92.69	96.59	97.92	98.49	98.67	98.73	95.69
0.3	0.4	0.3	**88.48**	92.93	96.55	97.90	98.42	98.60	98.67	95.94
0.3	0.3	0.4	88.15	**93.10**	96.79	98.08	98.57	98.75	98.81	**96.04**
0.4	0.2	0.4	87.37	92.97	**96.84**	**98.12**	98.62	98.80	98.87	95.94
0.4	0.0	0.6	85.26	92.24	96.61	98.06	98.62	**98.85**	98.88	95.50
0.4	0.1	0.5	86.27	92.66	96.70	98.11	**98.65**	98.84	**98.89**	95.73

accuracy significantly in low levels of noise condition, which achieves an absolute improvement of 0.42% in word accuracy when SNR = 20 dB compared to state-of-the-art method [19], but the combined loss (w/ combined loss) mainly obtains larger gains in strong noisy conditions (SNR $\leq$ 5 dB). This is expected since the combined loss makes the system more robust, and the audio-visual encoder helps resolve issues such as the effectiveness of combining the information of two modalities.

In particular, noticing that the result of using the decision fusion module (w/ A+V decision), the fusion likelihood is calculated from the audio network and visual network (without audio-visual network) which is used in [24], only gets better performance than baseline at -10 dB of babble noise. However, the decision fusion module (w/ AV decision), outperforms the baseline in both strong noisy conditions and low levels of noise condition due to the fact that it can explicitly model the reliability of each modality. The results show that each of the modifications including w/ AV encoder, w/ combined loss, and w/ AV decision can improve the word recognition accuracy of the AVSR baseline system under different SNR conditions. Furthermore, it can be seen that our hybrid fusion architecture can both benefit through their tight collaboration, which gets better performance than single feature fusion or

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:33:48 UTC from IEEE Xplore. Restrictions apply.



(a) Audio likelihood (b) Visual likelihood (c) Audio-Visual likelihood (d) Fusion likelihood

Fig. 4: Example results of output likelihoods, each row shows a typical case. The fusion likelihood is the final output of our model and gets the correct results. First row: audio network gets the incorrect word. Second row: visual network gets the incorrect word. Third row: audio-visual network gets the incorrect word. The true class label is marked with "Δ".

decision fusion strategy.

2) Impact of hyper-parameter values: Table III shows the word accuracy results of our method with different hyper-parameter values of the audio-visual decision module. Some representative hyper-parameters are given, which has the best performance under a certain SNR condition. It can be seen that with different hyper-parameter values, our proposed method obtained different performance at varying SNR levels of babble noise. Noticing that with $\alpha = 0.3$, $\beta = 0.4$ and $\gamma = 0.3$, the word accuracy gets best score at -10 dB while it gets worst performance in low levels of noise condition (SNR = 20 dB). Thus, we need to choose a set of suitable hyper-parameters to balance the word accuracy between the strong noisy conditions and low levels of noise condition. $\alpha = 0.3$, $\beta = 0.3$ and $\gamma = 0.4$ is used in our experiments, which has best average performance at varying SNR levels of babble noise.

E. Visualization

Fig. 4 visualizes three typical examples of audio, visual, audio-visual, and fusion likelihoods, where audio network, visual network, and audio-visual network get incorrect words, while the final fusion results are correct. The results show that our proposed hybrid fusion architecture is able to distinguish words with similar pronunciations, which can reliably handle the inherent ambiguity in both audio and visual modalities.

IV. Conclusions

In this work, we present a novel hybrid fusion based method for AVSR to address the challenge of inherent ambiguity in noisy environments. In order to find the internal correlation between spatial and temporal relationships from audio and visual information, a simple yet effective audio-visual encoder is proposed to capture more discriminative multi-modal feature from both modalities. A decision fusion module is designed in order to complementarily utilize the reliability measures of audio and visual information. Moreover, we introduce a combined loss to make the model learn the joint representation across audio-visual modalities robustly. By making full use of the information from different levels in a unified framework, the proposed hybrid fusion architecture can benefit through the tight collaboration between the feature fusion and decision fusion strategies, which is able to distinguish words with similar pronunciations and becomes robust to various noisy conditions. Experiments on LRW dataset demonstrate that our method achieves superior performance compared to other state-of-the-art methods in both clean and noisy conditions.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No. 61673030, U1613209), National Natural Science Foundation of Shenzhen (No. JCYJ20190808182209321).

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:33:48 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[2] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[3] Ara V Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy, "A coupled hmm for audio-visual speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, vol. 2, pp. II–2013.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[5] Jong-Seok Lee and Cheol Hoon Park, "Adaptive decision fusion for audio-visual speech recognition," *Speech recognition, technologies and applications*, pp. 275–296, 2008.

[6] Darryl Stewart, Rowan Seymour, Adrian Pass, and Ji Ming, "Robust audio-visual speech recognition under noisy audio-video conditions," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 175–184, 2013.

[7] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590–605, 2014.

[8] Pingping Wu, Hong Liu, Xiaofei Li, Ting Fan, and Xuewu Zhang, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 326–338, 2016.

[9] George Sterpu, Christian Saam, and Naomi Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proceedings of the International Conference on Multimodal Interaction*. ACM, 2018, pp. 111–115.

[10] Themos Stafylakis and Georgios Tzimiropoulos, "Zero-shot keyword spotting for visual speech recognition in-the-wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 513–529.

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

[12] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3444–3453.

[13] Stavros Petridis, Yujiang Wang, Zuwei Li, and Maja Pantic, "End-to-end audiovisual fusion with LSTMs," *arXiv preprint arXiv:1709.04343*, 2017.

[14] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia, "Modality attention for end-to-end audio-visual speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6565–6569.

[15] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to be published.

[16] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song, "Hearing lips: Improving lip reading by distilling speech recognizers," *arXiv preprint arXiv:1911.11502*, 2019.

[17] Md Rabiul Islam and Fayzur Rahman, "Hybrid feature and decision fusion based audio-visual speaker identification in challenging environment," *International Journal of Computer Applications*, vol. 9, no. 5, pp. 9–15, 2010.

[18] Shankar T Shivappa, Mohan Manubhai Trivedi, and Bhaskar D Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[19] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic, "End-to-end audiovisual speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6548–6552.

[20] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision (ACCV)*, 2016, pp. 87–103.

[21] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[22] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 513–520.

[23] Themos Stafylakis and Georgios Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

[24] Runwei Ding, Cheng Pang, and Hong Liu, "Audio-visual keyword spotting based on multidimensional convolutional neural network," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4138–4142.

[25] Themos Stafylakis, Muhammad Haris Khan, and Georgios Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and lstms," *Computer Vision and Image Understanding*, vol. 176, pp. 22–32, 2018.

[26] Chenhao Wang, "Multi-grained spatio-temporal modeling for lip-reading," *arXiv preprint arXiv:1908.11618*, 2019.

[27] Xing Zhao, Shuang Yang, Shiguang Shan, and Xilin Chen, "Mutual information maximization for effective lip reading," *arXiv preprint arXiv:2003.06439*, 2020.

[28] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:33:48 UTC from IEEE Xplore. Restrictions apply.