Recursive Joint Cross-Attention for Audio-Visual Speaker Verification

R. Gnana Praveen, Jahangir Alam
Computer Research Institute of Montreal, Montreal (Quebec) H3N 1M3, Canada
gnana-praveen.rajasekhar@crim.ca, jahangir.alam@crim.ca

Abstract

Speaker verification has been recently gaining a lot of attention using audio-visual fusion as faces and voices share close associations with each other. Though existing approaches based on audio-visual fusion showed improvement over unimodal systems, the potential of audio-visual fusion for speaker verification is not fully exploited. In this paper, we have investigated the prospect of effectively capturing both the intra- and inter-modal relationships across audio and visual modalities simultaneously, which can play a crucial role in significantly improving the fusion performance over unimodal systems. Specifically, we introduce a recursive fusion of the joint cross-attentional model, where a joint audio-visual feature representation is employed in the cross-attention framework in a recursive fashion in order to obtain more refined feature representations that can efficiently capture the intra- and inter-modal associations. Extensive experiments are conducted on the Voxceleb1 dataset to evaluate the proposed model. Results indicate that the proposed model is found to be promising in improving the performance of the audio-visual system.

1 Introduction

Speaker Verification (SV) deals with the problem of verifying the identity of a person, which has a wide range of applications in various fields such as forensics, commercial, and law enforcement applications (1). The task of SV has been predominantly explored using faces (2) and speech (3) signals independently. With the advancement of deep learning models, both face- and speech-based methods have individually achieved remarkable success (3). However, relying on individual modalities may often deteriorate the performance of the system when face or speech-based signals are degraded by extreme background noise or intra-variations such as pose, low illumination, manner of speaking, etc. Therefore, leveraging the fusion of both faces and voices has been gaining momentum as multiple modalities are often expected to complement each other (4). For instance, when speech modality is corrupted, we can rely on face to verify the identity of a person and vice-versa. Most of the existing audio-visual (A-V) fusion approaches for SV focused on score-level fusion (5; 6) or early feature-level fusion (7; 8). Though these methods have improved the fusion performance over unimodal systems, they fail to leverage the rich complementary inter-modal relationships among the audio and visual modalities.

In recent years, attention-based models have been explored to efficiently capture the complementary inter-modal associations across faces and voices (9; 10). Most of the existing attention-based models attempted to leverage the intra- and inter-modal relationships in a decoupled fashion. Another line of approaches focused on dealing with noisy modalities using a weighted combination of audio and visual modalities (4; 11). To effectively fuse audio and visual modalities it is very important to adeptly capture both intra- and inter-modal relationships. Intra-modal relationships offer rich information pertinent to the temporal dynamics of videos whereas inter-modal relationships provide significant information related to the complementarity of the modalities. Contrary to the prior approaches, we

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

have explored joint cross-attentional fusion in a recursive fashion to simultaneously capture both intra and inter-modal relationships to obtain robust A-V feature representations. Recursive attention has been previously explored successfully for emotion recognition (12) and event localization (13). By recursively fusing the features of audio and visual modalities we are able to achieve more refined feature representations in order to improve the performance of A-V fusion for SV. The major contributions of the proposed approach can be summarized as follows: (1) A recursive fusion of joint cross-attentional model is introduced to efficiently capture both intra- and inter-modal relationships across faces and voices. (2) The joint feature representation helps to mitigate heterogeneity issues, while simultaneously refining the feature vectors. (3) Extensive experiments are conducted on the voxceleb1 dataset to evaluate the robustness of the proposed model.

2 Related Work

The close association between faces and voices has attained much attention for the task of the cross-modal biometric matching system by projecting the features of individual modalities to a common representation space (14; 15). Sari et al. (16) explored a multi-view approach by transforming the individual feature representations into a common latent space and a shared classifier is used for both modalities for SV. Chen et al (17) leveraged the complementary information as a means of supervision to obtain robust A-V feature representations using a co-meta learning paradigm in a self-supervised learning framework. Tao et al (4) also explored the complementary relationship across audio and visual modalities to cleanse the noisy samples, where the consistency across the audio and visual modalities is used to discriminate the easy and hard samples. Another line of approaches is to deal with mitigating the impact of noisy modalities by leveraging complementary relationships. Shon et al (11) proposed an attention mechanism to assign higher attention scores to the modality exhibiting higher discrimination by leveraging the complementary nature across audio and visual modalities. Stefan et al (10) further extended the idea of (11) by introducing feature fusion of audio and visual modalities at intermediate layers to improve the quality of feature representations. Chen et al (7) explored the prospect of obtaining robust feature representations by investigating various fusion strategies at the embedding level and achieved the best performance using gating-based fusion. They further exploited data augmentation strategy to deal with extremely corrupted or missing modalities.

All the above-mentioned approaches fail to leverage the cross-modal interactions to effectively capture the rich inter-modal relationships. Cross-modal attention has been successfully explored in several applications such as weakly-supervised action localization (18), event localization (13), and emotion recognition (19). Recently, Bogdan et al (20) also explored cross-attention (CA) based on cross-correlation across the audio and visual modalities to effectively capture the complementary relationships. Meng et al (21) explored cross-modal attention by deploying cross-modal boosters in a pseudo-siamese structure to model one modality by exploiting the knowledge from another modality. However, these approaches focus only on inter-modal relationships (20) or capture the intra- and inter-modal relationships in a decoupled fashion (21). Praveen et al (22) explored a joint cross-attentional framework to jointly capture the intra and inter-modal relationships and showed better performance of SV. In this work, we further extend this approach by learning more robust A-V feature representations in a recursive fashion.

3 Recursive Joint Cross-Attention

Notations: Given an input video sub-sequence S, we uniformly sample L non-overlapping video segments and extract deep feature vectors from pre-trained models for audio and visual modalities. Let $\boldsymbol{X}_{\mathbf{a}}$ and $\boldsymbol{X}_{\mathbf{v}}$ denote the deep feature vectors of audio and visual modalities respectively for the given input video sub-sequence $\boldsymbol{S}$ of fixed size, which is expressed as $\boldsymbol{X}_{\mathbf{a}} = \{\boldsymbol{x}_{\mathbf{a}}^1, \boldsymbol{x}_{\mathbf{a}}^2, ..., \boldsymbol{x}_{\mathbf{a}}^L\} \in \mathbb{R}^{d_a \times L}$ and $\boldsymbol{X}_{\mathbf{v}} = \{\boldsymbol{x}_{\mathbf{v}}^1, \boldsymbol{x}_{\mathbf{v}}^2, ..., \boldsymbol{x}_{\mathbf{v}}^L\} \in \mathbb{R}^{d_v \times L}$ where d_a and d_v represent the dimensions of the audio and visual feature vectors, respectively, and $\boldsymbol{x}_{\mathbf{a}}^l$ and $\boldsymbol{x}_{\mathbf{v}}^l$ denotes the audio and visual feature vectors of the video segments, respectively, for $l = 1, 2, ..., L$ segments.

It has been shown that the performance of unified multimodal training may decline over that of individual modalities due to the differences in learning dynamics, noise topologies, etc. (23). Therefore, we have used fixed feature vectors of audio and visual modalities to train the proposed A-V fusion model. By deploying the joint feature representation in the CA framework in a recursive fashion, we are able to simultaneously enhance the intra- and inter-modeling of A-V relationships.

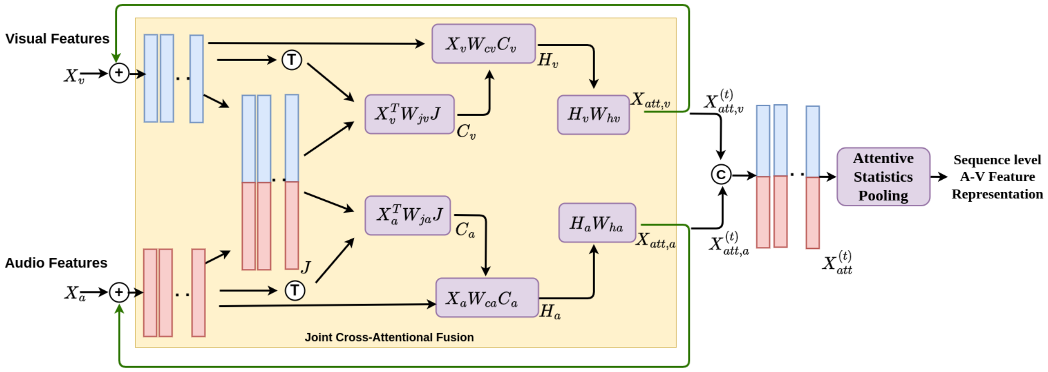

Figure 1: Block Diagram of the Recursive Joint Cross-Attention model for A-V fusion

The block diagram of the proposed approach is shown in Figure 1. The joint representation $\boldsymbol{J}$ is obtained by concatenating the audio and visual feature vectors as $\boldsymbol{J} = [\boldsymbol{X}_{\mathbf{a}}; \boldsymbol{X}_{\mathbf{v}}] \in \mathbb{R}^{d \times L}$ where $d = d_a + d_v$ denotes the feature dimension of concatenated features. The concatenated A-V feature representations ($\boldsymbol{J}$) of the given video sub-sequence ($\boldsymbol{S}$) are now employed in the CA framework to attend to the individual modalities. This helps to incorporate both intra- and inter-modal relationships in obtaining the attention weights of audio and visual modalities. Now the correlation across the joint feature representation and the individual modalities are obtained as a joint cross-correlation matrices, which is given by

$$\boldsymbol{C}_{\mathbf{a}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{a}}^{\top} \boldsymbol{W}_{\mathbf{ja}} \boldsymbol{J}}{\sqrt{d}}\right) \quad \text{and} \quad \boldsymbol{C}_{\mathbf{v}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{v}}^{\top} \boldsymbol{W}_{\mathbf{jv}} \boldsymbol{J}}{\sqrt{d}}\right) \tag{1}$$

where $\boldsymbol{W}_{\mathbf{ja}} \in \mathbb{R}^{d_a \times d}$, $\boldsymbol{W}_{\mathbf{jv}} \in \mathbb{R}^{d_v \times d}$ represents learnable weight matrices of audio and visual modalities respectively. The joint correlation matrices $\boldsymbol{C}_{\mathbf{a}}$ and $\boldsymbol{C}_{\mathbf{v}}$ for audio and visual modalities help to obtain the attention weights based on the semantic relevance of both across and within the modalities. The higher the correlation coefficient, the higher the correlation across the corresponding samples within the same modality as well as across the modality. Now the joint cross-correlation matrices are used to obtain the attention maps of audio and visual modalities, which are given by

$$\boldsymbol{H}_{\mathbf{a}} = ReLU(\boldsymbol{X}_{\mathbf{a}} \boldsymbol{W}_{\mathbf{ca}} \boldsymbol{C}_{\mathbf{a}}) \quad \text{and} \quad \boldsymbol{H}_{\mathbf{v}} = ReLU(\boldsymbol{X}_{\mathbf{v}} \boldsymbol{W}_{\mathbf{cv}} \boldsymbol{C}_{\mathbf{v}}) \tag{2}$$

where $\boldsymbol{W}_{\mathbf{ca}} \in \mathbb{R}^{L \times L}$,$\boldsymbol{W}_{\mathbf{cv}} \in \mathbb{R}^{L \times L}$ denote learnable matrices of audio and visual modalities respectively. These attention maps are used to obtain the attended features of audio and visual modalities as

$$\boldsymbol{X}_{\mathbf{att,a}} = \boldsymbol{H}_{\mathbf{a}} \boldsymbol{W}_{\mathbf{ha}} + \boldsymbol{X}_{\mathbf{a}} \quad \text{and} \quad \boldsymbol{X}_{\mathbf{att,v}} = \boldsymbol{H}_{\mathbf{v}} \boldsymbol{W}_{\mathbf{hv}} + \boldsymbol{X}_{\mathbf{v}} \tag{3}$$

where $\boldsymbol{W}_{\mathbf{ha}} \in \mathbb{R}^{L \times L}$ and $\boldsymbol{W}_{\mathbf{hv}} \in \mathbb{R}^{L \times L}$ denote the learnable weight matrices for audio and visual modalities respectively. In order to obtain more refined feature representations, the attended features are again fed as input to the joint cross-attentional model, which is given by

$$\boldsymbol{X}^{(t)}_{\mathbf{att,a}} = \boldsymbol{H}^{(t)}_{\mathbf{a}} \boldsymbol{W}^{(t)}_{\mathbf{ha}} + \boldsymbol{X}^{(t-1)}_{\mathbf{a}} \quad \text{and} \quad \boldsymbol{X}^{(t)}_{\mathbf{att,v}} = \boldsymbol{H}^{(t)}_{\mathbf{v}} \boldsymbol{W}^{(t)}_{\mathbf{hv}} + \boldsymbol{X}^{(t-1)}_{\mathbf{v}} \tag{4}$$

where $\boldsymbol{W}^{(t)}_{\mathbf{ha}} \in \mathbb{R}^{L \times L}$ and $\boldsymbol{W}^{(t)}_{\mathbf{hv}} \in \mathbb{R}^{L \times L}$ denote the learnable weight matrices of t^{th} iteration for audio and visual modalities respectively. Finally the attended features $\boldsymbol{X}^{(t)}_{\mathbf{att,a}}$ and $\boldsymbol{X}^{(t)}_{\mathbf{att,v}}$ obtained from the recursive fusion model are concatenated and fed to the attentive statistics pooling (ASP) to obtain sub-sequence or utterance-level representations. The utterance-level A-V feature representations are used to obtain the scores, where the additive angular margin softmax (AAMSoftmax) (24) loss function is used to optimize the parameters of the fusion model and ASP module.

4 Results and Discussion

Dataset: We have evaluated the proposed model on the Voxceleb1 dataset, obtained from YouTube videos under challenging environments. The dataset consists of 1,48,642 video clips, captured from

Table 1: Performance (EER) of the proposed approach by comparing to the existing fusion strategies (left) and comparison to state-of-the-art (right)

Fusion Method	**Validation Set**
Score-level Fusion	2.521
Feature Concatenation	2.489
Self-Attention	2.412
Cross-Attention	2.387
Joint Cross-Attention	2.125
RJCA (Ours)	1.946

Fusion Method	**Validation Set**	**Vox1-O Set**
Visual	3.720	3.779
Audio	2.553	2.529
Tao et al (4)	2.476	2.409
Praveen et al (22)	2.125	2.214
RJCA (Ours)	1.946	2.015

1,251 speakers, out of which 55% of the speakers are male. Each video clip has a duration of 4 to 145 seconds and the speakers are chosen to cover a diverse range of ethnicities, accents, professions, and ages. For our experiments, we have divided the voxceleb1 development set, which has 1211 speakers into training and validation sets. The training and validation splits were randomly selected as 1150 and 61 speakers respectively and reported our results on both the validation split and Vox1-O (Voxceleb1 original) test set for performance evaluation. It is worth mentioning that the model is trained only on the voxceleb1 dataset. Equal Error Rate (EER) is used as an evaluation metric, which refers to the point where the False Accept Rate (FAR) is equal to the False Reject Rate (FRR). A perfect scoring model should yield an EER of zero, so a lower EER value indicates a better performance.

Ablation Study: The results are reported based on the average of three runs for statistical stability. To obtain the audio and visual feature vectors, we have used Resnet-18 (25) for visual modality and ECAPA-TDNN (26) for audio modality similar to that of (22) to have a fair comparison. First, we have implemented a simple score-level fusion, where scores obtained from individual modalities are fused. Next, we explored feature concatenation, where the features of audio and visual modalities are concatenated and used to obtain the final score. We can observe that the fusion performance has been improved over simple score-level fusion. By employing a self-attention mechanism on the concatenated features of individual modalities, the fusion performance has been further improved by leveraging the intra-modal relationships. We also explored inter-modal relationships across the modalities using CA and found further improvement in fusion performance as shown in Table 1 (left). Now, we explored joint cross-attention, where joint feature representation is deployed in the CA framework to simultaneously capture both intra- and inter-modal relationships. Finally, we have introduced the recursive fusion of the attended features of individual modalities and observed that recursive fusion with 3 iterations helps in obtaining more refined feature representations and achieves the best performance among all the fusion strategies.

Comparison to state-of-the-art: Most of the existing methods used the Voxceleb2 development dataset for training the models for SV. However, we have used the Voxceleb1 dataset to validate the proposed approach and compared it with the recent state-of-the-art methods of SV using A-V fusion. Table 1 (right) shows the comparison of the proposed approach to recent state-of-the-art methods as well as individual modalities on both validation split of Voxceleb1 and Vox1-O datasets. First, we have conducted experiments with the individual modalities and found that audio performs relatively better than visual modality. In order to have a fair comparison, we have re-implemented the approach of (4) and (22) using the same experimental setup on the Voxceleb1 dataset. Tao et al (4) explored the complementary relationships as supervisory information to deal with noisy samples. Praveen et al (22) deployed the joint A-V representation in the CA framework and improved the fusion performance. Since the proposed approach employs recursive fusion with the joint cross-attentional framework to obtain robust feature representations, we can observe that the fusion performance has been further improved.

5 Conclusion

In this paper, we have presented a novel approach of recursive joint cross-attentional fusion for SV by effectively exploiting both inter- and intra-modal relationships across audio and visual modalities. Specifically, we have explored joint feature representation in the CA framework in a recursive fashion in order to obtain more refined A-V feature representations. The performance of the proposed approach can be further enhanced by training with the large-scale Voxceleb2 dataset as it can improve the generalization ability of the proposed approach.

Acknowledgments and Disclosure of Funding

The authors wish to acknowledge the funding from the Government of Canada's New Frontiers in Research Fund (NFRF) through grant NFRFR-2021-00338.

References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[3] Y. Tu, W. Lin, and M.-W. Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, vol. 10, pp. 99038–99049, 2022.

[4] R. Tao, K. A. Lee, Z. Shi, and H. Li, "Speaker recognition with two-step multi-modal deep cleansing," in *IEEE ICASSP*, pp. 1–5, 2023.

[5] Seyed, C. Greenberg, E. Singer, D. Olson, L. Mason, and J. Hernandez-Cordero, "The 2019 nist audio-visual speaker recognition evaluation," The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, -1, 2020-05-18 2020.

[6] J. Alam, G. Boulianne, L. Burget, M. Dahmane, M. S. Diez, O. Glembek, M. Lalonde, A. D. Lozano, P. Matějka, P. Mizera, L. Mošner, C. Noiseux, J. Monteiro, O. Novotný, O. Plchot, A. J. Rohdin, A. Silnova, J. Slavíček, T. Stafylakis, P.-L. St-Charles, S. Wang, and H. Zeinali, "Analysis of abc submission to nist sre 2019 cmn and vast challenge," in *Proceedings of Odyssey 2020 The Speaker and Language Recognition Workshop*, vol. 2020, pp. 289–295, International Speech Communication Association, 2020.

[7] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on voxceleb," in *Proc. Interspeech*, pp. 2252–2256, 2020.

[8] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1079–1092, 2021.

[9] P. Sun, S. Zhang, Z. Liu, Y. Yuan, T. Zhang, H. Zhang, and P. Hu, "A Method of Audio-Visual Person Verification by Mining Connections between Time Series," in *Proc. INTERSPEECH 2023*, pp. 3227–3231, 2023.

[10] S. Hörmann, A. Moiz, M. Knoche, and G. Rigoll, "Attention fusion for audio-visual person verification using multi-scale features," in *IEEE FG*, pp. 281–285, 2020.

[11] S. Shon, T.-H. Oh, and J. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *IEEE ICASSP*, pp. 3995–3999, 2019.

[12] R. G. Praveen, E. Granger, and P. Cardinal, "Recursive joint attention for audio-visual fusion in regression based emotion recognition," in *IEEE ICASSP*, pp. 1–5, 2023.

[13] B. Duan, H. Tang, W. Wang, Z. Zong, G. Yang, and Y. Yan, "Audio-visual event localization via recursive fusion by joint co-attention," in *IEEE WACV*, pp. 4012–4021, 2021.

[14] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *IEEE CVPR*, 2018.

[15] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. of ECCV*, 2018.

[16] L. Sarı, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *IEEE ICASSP*, pp. 6194–6198, 2021.

[17] H. Chen, H. Zhang, L. Wang, K. A. Lee, M. Liu, and J. Dang, "Self-supervised audio-visual speaker representation with co-meta learning," in *IEEE ICASSP*, pp. 1–5, 2023.

[18] J.-T. Lee, M. Jain, H. Park, and S. Yun, "Cross-attentional audio-visual fusion for weakly-supervised action localization," in *Proc. of the ICLR*, 2021.

[19] R. G. Praveen, E. Granger, and P. Cardinal, "Cross attentional audio-visual fusion for dimensional emotion recognition," in *IEEE FG*, pp. 1–8, 2021.

[20] B. Mocanu and T. Ruxandra, "Active speaker recognition using cross attention audio-video fusion," in *Proc. of the EUVIP*, pp. 1–6, 2022.

[21] M. Liu, K. A. Lee, L. Wang, H. Zhang, C. Zeng, and J. Dang, "Cross-modal audio-visual co-learning for text-independent speaker verification," in *IEEE ICASSP*, pp. 1–5, 2023.

[22] R. G. Praveen and J. Alam, "Audio-visual speaker verification via joint cross-attention," 2023.

[23] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?," in *CVPR*, pp. 12692–12702, 2020.

[24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on CVPR*, pp. 4685–4694, 2019.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.

[26] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, pp. 3830–3834, 2020.

A Appendix

A.1 Ablation study with varying number of recursions

In order to understand the impact of the recursive fusion, we have performed experiments by varying the number of iterations and obtained the best performance at 3 iterations as shown in Table 2. Beyond that, we observe a decline in the fusion performance, which can be attributed to the fact that the recursion acts as a regularizer and improves the generalization ability of the proposed model. The results are reported based on the average of three runs for statistical stability.

Table 2: Performance (EER) of the proposed approach by varying the number of recursions

Fusion Method	**Validation Set**
RJCA Fusion (t = 2)	2.029
RJCA Fusion (t = 3)	1.946
RJCA Fusion (t = 4)	1.995
RJCA Fusion (t = 5)	2.159