Audio-Visual Speaker Verification via Joint Cross-Attention

Gnana Praveen Rajasekhar(✉) and Jahangir Alam(✉)

Computer Research Institute of Montreal, Montreal, QC H3N 1M3, Canada
{gnana-praveen.rajasekhar,jahangir.alam}@crim.ca

Abstract. Speaker verification has been widely explored using speech signals, which has shown significant improvement using deep models. Recently, there has been a surge in exploring faces and voices as they can offer more complementary and comprehensive information than relying only on a single modality of speech signals. Though current methods in the literature on the fusion of faces and voices have shown improvement over that of individual face or voice modalities, the potential of audio-visual fusion is not fully explored for speaker verification. Most of the existing methods based on audio-visual fusion either rely on score-level fusion or simple feature concatenation. In this work, we have explored cross-modal joint attention to fully leverage the inter-modal complementary information and the intra-modal information for speaker verification. Specifically, we estimate the cross-attention weights based on the correlation between the joint feature presentation and that of the individual feature representations in order to effectively capture both intra-modal as well inter-modal relationships among the faces and voices. We have shown that efficiently leveraging the intra- and inter-modal relationships significantly improves the performance of audio-visual fusion for speaker verification. The performance of the proposed approach has been evaluated on the Voxceleb1 dataset. Results show that the proposed approach can significantly outperform the state-of-the-art methods of audio-visual fusion for speaker verification.

Keywords: Cross-attention · Audio-visual fusion · Speaker verification · Joint-attention

1 Introduction

Speaker verification is the task of verifying the identity of a person, which is primarily carried out using acoustic samples. It has become a key technology for person authentication in various real-world applications such as customer authentication, security applications, etc [14,19]. In recent years, the performance of speaker verification has been significantly boosted using deep learning models based on acoustic samples such as x-vector [41], xi-vector [20], and ECAPA-TDNN [9]. However, in a noisy acoustic environment, it would be difficult to distinguish different speakers only based on speech signals. Therefore,

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 18–31, 2023.
https://doi.org/10.1007/978-3-031-48312-7_2

other modalities such as face, iris, and fingerprints are also explored for verifying the person's identity. Out of all the modalities, face and voice share a very close association with each other in identifying a person's identity [16]. Authenticating the identity of a person from videos has been widely explored in the literature by relying either on faces [15,32,44] or voices [2,40,59]. Inspired by the close association between faces and voices, audio-visual (A-V) systems [6,52,55,58] have been proposed for speaker verification. However, effectively leveraging the fusion of voices and faces for speaker verification is not fully explored in the literature [22,46]. Face and voice provide diverse and complementary relationships with each other, which plays a key role in outperforming the performance of individual modalities.

Conventionally, A-V fusion can be achieved by three major fusion strategies: feature-level fusion, model-level fusion, and decision-level fusion [54]. Feature-level fusion (or early fusion) is performed by naively concatenating the features of individual audio and visual modalities, which is further used for predicting the final outputs. Model-level fusion deals with specialized architectures for fusion based on models such as deep networks [56], Hidden Markov Model (HMM) [57], and kernel methods [4]. In decision-level fusion, audio and visual modalities are trained independently end-to-end, and then the scores obtained from the individual modalities are fused to obtain the final scores. It requires little training and is easy to implement, however, it neglects the interactions across the modalities and thereby shows limited improvement over the individual performances of faces and voices. Though feature (or early-level) fusion allows the audio and visual modalities to interact with each other at the feature level, they fail to effectively capture the complementary inter-modal and intra-modal relationships with each other. Most of the existing approaches for speaker verification based on A-V fusion either fall in the category of decision-level fusion, where fusion is performed at score level, or early feature-level fusion, which relies on early feature concatenation of audio and visual modalities. Even though naive feature concatenation or using score level fusion shows improvement in the performance of speaker verification, it does not fully leverage the intra-modal and inter-modal relationships among the audio and visual modalities. In some of the videos, the voices might be corrupted due to background clutter. On the other hand, face images can also be corrupted due to several factors such as occlusion, pose, poor resolution, etc. Intuitively, an ideal strategy of A-V fusion should give more importance to the modality, exhibiting better-discriminating features by fully exploiting the complementary relationships with each other.

Recently, attention mechanisms have been explored to focus on the more relevant modalities of the video clips by assigning higher attention weights to the modality exhibiting higher discrimination among the speakers [38]. In this work, we have investigated the prospect of leveraging the complementary relationships among the faces and voices, while still leveraging the intra-modal temporal dynamics within the same modality to improve the performance of the system than that of individual audio and visual modalities. Specifically, a joint feature representation is introduced to the joint cross-attentional fusion model

along with the feature representations of individual modalities to simultaneously capture both the intra-modal relationships and complementary inter-modal relationships. The major contributions of this paper are as follows:

- A joint cross-attentional model is explored for an effective fusion of faces (visual) and voices (audio) by leveraging both the intra-modal and inter-modal relationships for speaker verification.
- Deploying the joint feature representation also helps to reduce the heterogeneity among the audio and visual features, thereby resulting in better A-V feature representations
- A detailed set of experiments are conducted to show that the proposed approach is able to outperform the state-of-the-art A-V fusion models for speaker verification.

2 Related Work

2.1 Audio-Visual Fusion for Speaker Verification

Nagrani et al. [27] is one of the early works to investigate the close association of voices and faces and proposed a cross-modal biometric matching system. They have attempted to match a given static face or dynamics video with the corresponding voice and vice-versa. They have further explored joint embeddings for the task of person verification, where the idea is to detect whether the faces and voices come from the same video or not [26]. Wen et al. [53] also explored shared representation space for voices and faces and presented a disjoint mapping network for cross-modal biometric matching by mapping the modalities individually to their common covariates. Tao et al. [45] proposed a cross-modal discriminative network based on the faces and voices of a given video. They have also investigated the association of faces and voices, whether the faces and voices come from the same person or not, and their application for speaker recognition. Another interesting work on cross-modal speaker verification was done by Nawaz et al. [30], where they analyzed the impact of languages for cross-modal biometric matching tasks in the wild. They have shown that both face and speaker verification systems rely on spoken languages, which is caused due to the domain shift across different languages. Leda et al. [37] attempted to leverage the complementary information of audio and visual modalities for speaker verification using a multi-view model, which uses a shared classifier to map audio and visual into the same space. Wang [50] explored various fusion strategies at the feature level and decision level, and showed that high-level features of audio and visual modalities share more semantic information than low-level features, which helps in improving the performance of the system. Chen et al. [3] proposed a co-meta learning paradigm for learning A-V feature representations in a self-supervised learning framework. In particular, they have leveraged the complementary information among the audio and visual modalities as a means of supervisory signal to obtain robust A-V feature representations. Meng et al. [22] also proposed a co-learning cross-modal framework, where the features of

each modality are obtained by exploiting the knowledge from another modality using cross-modal boosters in a pseudo-siamese structure. Tao et al. [46] proposed a two-step A-V deep cleansing framework to deal with the noisy samples. They have used audio modality to discriminate the easy and complex samples as a coarse-grained cleansing, which is further refined as a fine-grained cleansing using the visual modality. Unlike prior approaches, we have investigated the prospect of leveraging attention mechanisms to fully exploit the complementary inter-modal and intra-modal relationships among the audio and visual modalities for speaker verification.

2.2 Attention Models for Audio-Visual Fusion

Attention mechanisms are widely used in the context of multimodal fusion with various modalities such as audio and text [21,25], visual and text [23,51], etc. Stefan et al. [13] proposed a multi-scale feature fusion approach to obtain robust A-V feature representations. They have fused the features at intermediate layers of the audio and visual backbones, which are finally combined with the feature vectors of individual modalities in the shared common space to obtain the final A-V feature representations. Peiwen et al. [43] proposed a novel fusion strategy, that involves weight-enhanced attentive statistics pooling for both modalities, which exhibit a strong correlation with each other. They further obtain keyframes in both modalities using cycle consistency loss along with a gated attention mechanism to obtain robust A-V embeddings for speaker verification. Shon et al. [38] explored an attention mechanism to conditionally select the relevant modality in order to deal with noisy modalities. They have leveraged the complementary information among the audio and visual modalities by assigning higher attention weights to the modality, exhibiting higher discrimination for speaker verification. Chen et al. [5] investigated various fusion strategies and loss functions to obtain robust A-V feature representations for speaker verification. They have further evaluated the impact of the fusion strategies on extremely missing or corrupted modalities by leveraging the data augmentation strategy to discriminate the noisy and clean embeddings. Cross-modal attention among the audio and visual modalities has been successfully explored in several applications such as weakly-supervised action localization [18], A-V event localization [10], and emotion recognition [34,36]. Bogdan et al. [24] explored a cross-attention mechanism for the A-V fusion based on cross-correlation across the audio and visual modalities. The features of each modality are learned under the constraints of other modalities. However, they focus only on inter-modal relationships and fail to exploit the intra-modal relationships. Praveen et al. [33] explored a joint cross-attentional (JCA) framework for dimensional emotion recognition, which is closely related to our work. However, we have further adapted the JCA model for speaker verification by introducing the attentive statistics pooling module.

3 Problem Formulation

For an input video sub-sequence S, L non-overlapping video segments are uniformly sampled, and the corresponding deep feature vectors are obtained from the pre-trained models of audio and visual modalities. Let $\boldsymbol{Z_a}$ and $\boldsymbol{Z_v}$ 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:

$$\boldsymbol{Z_a} = \{\boldsymbol{z_a^1}, \boldsymbol{z_a^2}, ..., \boldsymbol{z_a^L}\} \in \mathbb{R}^{d_a \times L} \tag{1}$$

$$\boldsymbol{Z_v} = \{\boldsymbol{z_v^1}, \boldsymbol{z_v^2}, ..., \boldsymbol{z_v^L}\} \in \mathbb{R}^{d_v \times L} \tag{2}$$

where d_a and d_v represent the dimensions of the audio and visual feature vectors, respectively, and $\boldsymbol{z_a^l}$ and $\boldsymbol{z_v^l}$ denotes the audio and visual feature vectors of the video segments, respectively, for $l = 1, 2, ..., L$ segments The objective of the problem is to estimate the speaker verification model $f : \boldsymbol{Z} \rightarrow \boldsymbol{Y}$ from the training data $\boldsymbol{Z}$, where $\boldsymbol{Z}$ denotes the set of audio and visual feature vectors of the input video segments and $\boldsymbol{Y}$ represents the speaker identity of the corresponding video sub-sequence S.

4 Proposed Approach

4.1 Visual Network

Faces from videos involve both appearance and temporal dynamics of video sequences, which can provide information pertaining to a wide range of intra-variations of visual modality. Effectively capturing the spatiotemporal dynamics of facial videos plays a key role in obtaining robust feature representations. Long Short-Term Memory Networks (LSTMs) have been found to be promising in modeling the long-term temporal cues in sequence representations for various applications [35,48]. In this work, we have used Resnet18 [12] trained on the Voxceleb1 dataset [28] to obtain the spatial feature representations of the video frames. Conventionally, the size of the visual feature vectors of the last convolutional layer will be $512 \times 7 \times 7$, which is fed to the pooling layer to reduce the spatial dimension from 7 to size 1. However, this spatial reduction may leave out some useful information, which may deteriorate the performance of the system. Therefore, as suggested by [10], we have deployed scaled dot-product of audio and visual feature vectors for each segment in order to leverage the audio feature vectors to smoothly reduce the spatial dimensions of video feature vectors. Then, we encode the temporal dynamics of the segments of the sequence of visual feature vectors using Bi-directional LSTM with residual embedding. Finally, the obtained feature vectors of visual modality are stacked to form a matrix of visual feature vectors as shown by

$$\boldsymbol{X_v} = (\boldsymbol{x_v^1}, \boldsymbol{x_v^2}, ..., \boldsymbol{x_v^L}) \in \mathbb{R}^{d_v \times L} \tag{3}$$

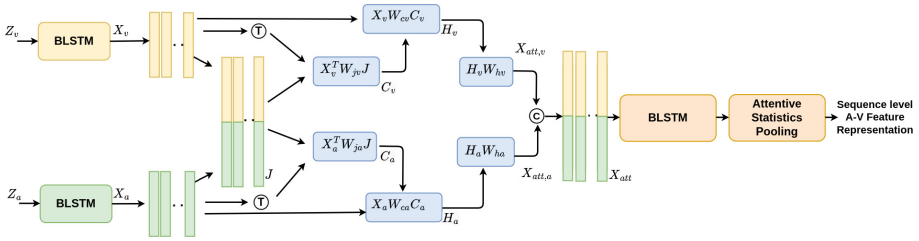

Fig. 1. Block Diagram of the Joint cross-attention model for A-V fusion.

4.2 Audio Network

With the advent of deep neural networks, speaker verification based on deep feature vectors has shown significant improvement over the conventional i-vector [7] based methods. One of the most widely used deep feature vector embeddings is the x-vector paradigm [41], which uses time-delay neural network (TDNN) and statistics pooling. Several variants of TDNN such as Extended TDNN (ETDNN) [42] and Factored TDNN (FTDNN) [47] have been introduced to boost the performance of the system. Recently, ECAPA-TDNN [9] has been introduced for speaker verification, which has shown significant improvement by leveraging the residual and squeeze-and-excitation (SE) components. So we have also explored ECAPA-TDNN to obtain the deep feature vectors of the audio segments. In order to exploit the temporal dynamics in the speech sequence, LSTMs have also been explored for speaker embedding extraction [1,60]. Similar to that of visual modality, we have also used Bi-directional LSTMs with residual embedding to encode the obtained audio feature vectors. Finally, the audio feature vectors of L video clips are stacked to obtain a matrix, shown as

$$\boldsymbol{X}_{\mathbf{a}} = (\boldsymbol{x}_{\mathbf{a}}^{1}, \boldsymbol{x}_{\mathbf{a}}^{2}, ..., \boldsymbol{x}_{\mathbf{a}}^{L}) \in \mathbb{R}^{d_a \times L} \tag{4}$$

4.3 Joint Cross-Attentional AV-Fusion

Though audio-visual fusion can be achieved through unified multimodal training, it was found that multimodal performance often declines over that of individual modalities [49]. This has been attributed to a number of factors, such as differences in learning dynamics for audio and visual modalities [49], different noise topologies, with some modality streams containing more or less information for the task at hand, as well as specialized input representations [29]. Therefore, we have obtained deep feature vectors for the individual audio and visual modalities independently, which are then fed to the joint cross-attentional module for audio-visual fusion.

Since multiple modalities convey more diverse information than a single modality, effectively leveraging the intra-modal and inter-modal complementary relationships among the audio and visual modalities plays a key role in efficient audio-visual fusion. In this work, we have explored joint cross-attentional

fusion to encode the intra-modal and inter-modal relationships simultaneously in a joint framework. Specifically, the joint A-V feature representation, obtained by concatenating the audio and visual features is also fed to the fusion module along with the feature representations of individual modalities. By deploying the joint representation, features of each modality attend to themselves, as well as other modalities, thereby simultaneously capturing the semantic inter-modal and intra-modal relationships among audio and visual modalities. Leveraging the joint representation also helps in reducing the heterogeneity among the audio and visual modalities, which further improves the performance of speaker verification. A block diagram of the proposed model is shown in Fig. 1. The joint representation of audio-visual features, $\boldsymbol{J}$, is obtained by concatenating the audio and visual feature vectors:

$$\boldsymbol{J} = [\boldsymbol{X}_{\mathbf{a}}; \boldsymbol{X}_{\mathbf{v}}] \in \mathbb{R}^{d \times L} \tag{5}$$

where $d = d_a + d_v$ denotes the feature dimension of concatenated features.

The concatenated audio-visual feature representations ($\boldsymbol{J}$) of the given video sub-sequence ($\boldsymbol{S}$) are now used to attend to the feature representations of individual modalities $\boldsymbol{X}_{\mathbf{a}}$ and $\boldsymbol{X}_{\mathbf{v}}$. The joint correlation matrix $\boldsymbol{C}_{\mathbf{a}}$ across the audio features $\boldsymbol{X}_{\mathbf{a}}$, and the combined audio-visual features $\boldsymbol{J}$ are given by:

$$\boldsymbol{C}_{\mathbf{a}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{a}}^{T}\boldsymbol{W}_{\mathbf{ja}}\boldsymbol{J}}{\sqrt{d}}\right) \tag{6}$$

where $\boldsymbol{W}_{\mathbf{ja}} \in \mathbb{R}^{L \times L}$ represents learnable weight matrix across the audio and combined audio-visual features, and T denotes transpose operation. Similarly, the joint correlation matrix for visual features is given by:

$$\boldsymbol{C}_{\mathbf{v}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{v}}^{T}\boldsymbol{W}_{\mathbf{jv}}\boldsymbol{J}}{\sqrt{d}}\right) \tag{7}$$

The joint correlation matrices $\boldsymbol{C}_{\mathbf{a}}$ and $\boldsymbol{C}_{\mathbf{v}}$ for audio and visual modalities provide a semantic measure of relevance not only across the modalities but also within the same modality. A higher correlation coefficient of the joint correlation matrices $\boldsymbol{C}_{\mathbf{a}}$ and $\boldsymbol{C}_{\mathbf{v}}$ shows that the corresponding samples are strongly correlated within the same modality as well as other modality. Therefore, the proposed approach is able to efficiently leverage the complementary nature of audio and visual modalities (i.e., inter-modal relationship) as well as intra-modal relationships, thereby improving the performance of the system. After computing the joint correlation matrices, the attention weights of audio and visual modalities are estimated.

For the audio modality, the joint correlation matrix $\boldsymbol{C}_{\mathbf{a}}$ and the corresponding audio features $\boldsymbol{X}_{\mathbf{a}}$ are combined using the learnable weight matrices $\boldsymbol{W}_{\mathbf{ca}}$ to compute the attention weights of audio modality, which is given by

$$\boldsymbol{H}_{\mathbf{a}} = ReLU(\boldsymbol{X}_{\mathbf{a}}\boldsymbol{W}_{\mathbf{ca}}\boldsymbol{C}_{\mathbf{a}}) \tag{8}$$

where $\boldsymbol{W}_{\mathbf{ca}} \in \mathbb{R}^{d_a \times d_a}$ and $\boldsymbol{H}_{\mathbf{a}}$ represents the attention maps of the audio modality.

Similarly, the attention maps ($\boldsymbol{H}_{\mathbf{v}}$) of visual modality are obtained as

$$\boldsymbol{H}_{\mathbf{v}} = ReLU(\boldsymbol{X}_{\mathbf{v}} \boldsymbol{W}_{\mathbf{cv}} \boldsymbol{C}_{\mathbf{v}}) \tag{9}$$

where $\boldsymbol{W}_{\mathbf{cv}} \in \mathbb{R}^{d_v \times d_v}$ denote the learnable weight matrices.

Then, the attention maps are used to compute the attended features of audio and visual modalities as:

$$\boldsymbol{X}_{\mathbf{att,a}} = \boldsymbol{H}_{\mathbf{a}} \boldsymbol{W}_{\mathbf{ha}} + \boldsymbol{X}_{\mathbf{a}} \tag{10}$$

$$\boldsymbol{X}_{\mathbf{att,v}} = \boldsymbol{H}_{\mathbf{v}} \boldsymbol{W}_{\mathbf{hv}} + \boldsymbol{X}_{\mathbf{v}} \tag{11}$$

where $\boldsymbol{W}_{\mathbf{ha}} \in \mathbb{R}^{d \times d_a}$ and $\boldsymbol{W}_{\mathbf{hv}} \in \mathbb{R}^{d \times d_v}$ denote the learnable weight matrices for audio and visual modalities respectively.

The attended audio and visual features, $\boldsymbol{X}_{\mathbf{att,a}}$ and $\boldsymbol{X}_{\mathbf{att,v}}$ are further concatenated to obtain the A-V feature representation, which is given by:

$$\widehat{\mathbf{X}} = [\boldsymbol{X}_{\mathbf{att,v}}; \boldsymbol{X}_{\mathbf{att,a}}] \tag{12}$$

The attended audio-visual feature vectors are fed to the Bi-directional LSTM in order to capture the temporal dynamics of the attended joint audio-visual feature representations. The segment-level audio-visual feature representations are in turn fed to the attentive statistics pooling (ASP) [31] in order to obtain the subsequence or utterance-level representation of the audio-visual feature vectors. Finally, the embeddings of the final audio-visual feature representations are used to obtain the scores, where the additive angular margin softmax (AAMSoftmax) [8] loss function is used to optimize the parameters of the fusion model and ASP module.

5 Experimental Methodology

5.1 Datasets

The proposed approach has been evaluated on the VoxCeleb1 dataset [28], obtained from videos of YouTube interviews, captured in a large number of challenging multi-speaker acoustic environments. The dataset contains 1,48,642 video clips from 1,251 speakers, which is gender-balanced with 55% of the speakers being male. The speakers are selected from a wide range of different ethnicities, accents, professions, and ages. The duration of the video clips ranges from 4 s to 145 s. In our experimental framework, we split the voxceleb1 development set (comprised of videos from 1211 speakers) into training and validation sets. We have randomly selected 1150 speakers for training and 61 speakers for validation. We have also reported our results on the Vox1-O (Voxceleb1 Original) test set for performance evaluation. This test set consists of 37720 trials from 40 speakers.

5.2 Evaluation Metric

In order to evaluate the performance of our proposed approach, we used equal error rate (EER) as an evaluation metric, which has been widely used for speaker verification in the literature [7,24]. It depicts the error rate when the False Accept Rate (FAR) is equal to the False Reject Rate (FRR). So the lower the EER, the higher the reliability of the system.

5.3 Implementation Details

For the visual modality, the facial images are taken from the images provided by the organizers of the dataset. For regularizing the network, dropout is used with $p = 0.8$ on the linear layers. The initial learning rate of the network was set to be $1e-2$ is used for the Adam optimizer. Also, weight decay of $5e-4$ is used. The batch size of the network is set to 400. Data augmentation is performed on the training data by random cropping, which produces a scale-invariant model. The number of epochs is set to be 50 and early stopping is used to obtain the best weights of the network.

For training the audio network, 80-dimensional Mel-FilterBank (MFB) features are extracted using an analysis window size of 25 ms over a frameshift of 10 ms. The acoustic features are randomly augmented on-the-fly with either MUSAN noise, speed perturbation with a rate between 0.95 and 1.05, or reverberation [39]. In addition, we use SpecAugment [17] for applying frequency and time masking on the MFB features. The initial weights of the audio network are initialized with values from the normal distribution and the network is trained for a maximum of 100 epochs, and early stopping is used. The network is optimized using Adam optimizer with the initial learning rate of 0.001 and the batch size is fixed to be 400. In order to prevent the network from over-fitting, dropout is used with p = 0.5 after the last linear layer. Also, weight decay of $5e-4$ is used for all the experiments.

For the fusion network, we used hyperbolic tangent functions for the activation of cross-attention modules. The dimension of the extracted features of audio modality is set to 192 and visual modality as 512. In the joint cross-attention module, the initial weights of the joint cross-attention matrix are initialized with the Xavier method [11] and the weights are updated using the Adam optimizer. The initial learning rate is set to be 0.001 and batch size is fixed to be 100. Also, a dropout of 0.5 is applied on the attended A-V features and weight decay of $5e-4$ is used for all the experiments.

6 Results and Discussion

6.1 Ablation Study

In order to analyze the performance of the proposed fusion model, we compare the proposed fusion model with some of the widely-used fusion strategies for speaker verification. One of the widely used fusion strategies is score-level fusion,

where the scores of the individual modalities are obtained and fused together to estimate the identity of a person. Another common approach for A-V fusion is based on early fusion, where the deep features of audio and visual modalities are concatenated immediately after being extracted, and the concatenated version of the individual modalities is used to obtain the final scores. As we can observe in the Table, the proposed fusion model consistently outperforms both the early fusion and the score level (decision level) by leveraging the semantic intra-modal and inter-modal relationships among the audio and visual modalities for speaker verification.

In order to analyze the contribution of the LSTMs in improving the modeling of intra-modal relationships for both individual feature representations and the final attended A-V feature representations, we have carried out a series of experiments with and without Bi-directional LSTMs (BLSTM). The experimental results to analyze the impact of BLSTMs have been shown in Table 1 Initially, we conducted an experiment without using Bi-LSTMs with the proposed fusion model. Then, we introduced Bi-LSTMs only for modeling the temporal dynamics of individual feature representations. We can observe that the performance of the proposed fusion model with the U-BLSTMs for individual feature representations has been improved. Now, we introduce BLSTMs for modeling the temporal dynamics of the final A-V attended feature representations. As observed in Table 1, the performance of the proposed fusion model has been further improved by introducing J-BLSTMs for modeling the temporal dynamics of final A-V feature representations.

Table 1. Performance of various fusion strategies on the validation set.

Fusion Method	EER
Feature Concatenation (Early Fusion)	2.489
Score-level Fusion (Decision-level)	2.521
Proposed Fusion (JCA) without BLSTMs	2.315
Proposed Fusion (JCA) with U-BLSTMs	2.209
Proposed Fusion (JCA) with U-BLSTMs and J-BLSTMs	2.173

6.2 Comparision to State-of-the-Art

In order to compare with state-of-the-art, we have used the recently proposed A-V fusion model based on two-step multimodal deep cleansing [46]. We have used their deep cleansing approach as a baseline and extended their approach by introducing our proposed fusion model to obtain robust A-V feature representations. The experimental results of the proposed approach in comparison to that of [46] are shown in Table 2. We have reported the results for both the validation set and the Vox1-O test partition of the Voxceleb1 dataset. In order

to analyze the fusion performance of the proposed model, we have also reported the results for the individual audio and visual modalities. We can observe that the proposed fusion model clearly outperforms the performance of individual modalities. We can also observe that by introducing the proposed fusion model, the performance of the system has been improved better than that of [46].

Table 2. Performance of the proposed approach in comparison to state-of-the-art on the validation set and Vox1-O set.

Fusion Method	Validation Set	Vox1-O Set
Face	3.720	3.779
Speech	2.553	2.529
Tao et al. [46]	2.476	2.4096
Proposed Fusion Model	2.125	2.214

7 Conclusion

In this paper, we present a joint cross-attentional A-V fusion model for speaker verification in videos. Unlike prior approaches, we effectively leverage the intra-modal and complementary inter-modal relationships among the audio and visual modalities. In particular, we obtain the deep features of audio and visual modalities from pre-trained networks, which are fed to the fusion model along with the joint representation. Then semantic relationships among audio and visual modalities are obtained based on the cross-correlation between the individual feature representations and the joint A-V feature representation (concatenated version of audio and visual features). The attention weights obtained from the cross-correlation matrix are used to estimate the attended feature vectors of audio and visual modalities. The modeling of intra-modal relationships in the proposed system has been further improved by leveraging Bi-directional LSTMs to model the temporal dynamics of both the individual feature representations and the final attended A-V feature representations. Experiments have shown that the proposed approach outperforms the state-of-the-art approaches for speaker verification.

Acknowledgments. 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. Alam, J., Fathan, A., Kang, W.H.: Text-independent speaker verification employing CNN-LSTM-TDNN hybrid networks. In: Speech and Computer, pp. 1–13 (2021)
2. Alam, J., Kang, W.H., Fathan, A.: Hybrid neural network with cross- and self-module attention pooling for text-independent speaker verification. In: IEEE ICASSP, pp. 1–5 (2023)
3. Chen, H., Zhang, H., Wang, L., Lee, K.A., Liu, M., Dang, J.: Self-supervised audio-visual speaker representation with co-meta learning. In: IEEE ICASSP, pp. 1–5 (2023)
4. Chen, J., Chen, Z., Chi, Z., Fu, H.: Emotion recognition in the wild with feature fusion and multiple kernel learning. In: ICMI, pp. 508–513 (2014)
5. Chen, Z., Wang, S., Qian, Y.: Multi-modality matters: a performance leap on VoxCeleb. In: Proceedings of Interspeech, pp. 2252–2256 (2020)
6. Chetty, G., Wagner, M.: Audiovisual speaker identity verification based on cross modal fusion. In: Proceedings of Auditory-Visual Speech Processing, p. paper P37 (2007)
7. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE TASLP **19**(4), 788–798 (2011)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: IEEE/CVF Conference on CVPR, pp. 4685–4694 (2019)
9. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: Proceedings of Interspeech, pp. 3830–3834 (2020)
10. Duan, B., Tang, H., Wang, W., Zong, Z., Yang, G., Yan, Y.: Audio-visual event localization via recursive fusion by joint co-attention. In: IEEE WACV, pp. 4012–4021 (2021)
11. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: ACAIS. vol. 9, pp. 249–256 (2010)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
13. Hörmann, S., Moiz, A., Knoche, M., Rigoll, G.: Attention fusion for audio-visual person verification using multi-scale features. In: IEEE FG, pp. 281–285 (2020)
14. Jelil, S., Shrivastava, A., Das, R.K., Prasanna, S.R.M., Sinha, R.: Speechmarker: a voice based multi-level attendance application. In: Interspeech, pp. 3665–3666 (2019)
15. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: IEEE Conference on CVPR, pp. 4873–4882 (2016)
16. Kim, C., Shin, H.V., Oh, T.H., Kaspar, A., Elgharib, M., Matusik, W.: On learning associations of faces and voices. In: Proceedings of the ACCV (2018)
17. Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S.: A study on data augmentation of reverberant speech for robust speech recognition. In: IEEE ICASSP, pp. 5220–5224 (2017). https://doi.org/10.1109/ICASSP.2017.7953152
18. Lee, J.T., Jain, M., Park, H., Yun, S.: Cross-attentional audio-visual fusion for weakly-supervised action localization. In: Proceedings of the ICLR (2021)
19. Lee, K., Larcher, A., Thai, H., Ma, B., Li, H.: Joint application of speech and speaker recognition for automation and security in smart home. In: INTERSPEECH, pp. 3317–3318 (2011)

20. Lee, K.A., Wang, Q., Koshinaka, T.: Xi-vector embedding for speaker recognition. IEEE SPL **28**, 1385–1389 (2021)
21. Lee, Y., Yoon, S., Jung, K.: Multimodal speech emotion recognition using cross attention with aligned audio and text. In: INTERSPEECH, pp. 2717–2721 (2020)
22. Liu, M., Lee, K.A., Wang, L., Zhang, H., Zeng, C., Dang, J.: Cross-modal audio-visual co-learning for text-independent speaker verification. In: IEEE ICASSP, pp. 1–5 (2023)
23. Ma, C., Shen, C., Dick, A., Wu, Q., Wang, P., Hengel, A.v.d., Reid, I.: Visual question answering with memory-augmented networks. In: CVPR, pp. 6975–6984 (2018)
24. Mocanu, B., Ruxandra, T.: Active speaker recognition using cross attention audio-video fusion. In: Proceedings of the EUVIP, pp. 1–6 (2022)
25. N., K.D., Patil, A.: Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In: INTERSPEECH, pp. 4243–4247 (2020)
26. Nagrani, A., Albanie, S., Zisserman, A.: Learnable pins: cross-modal embeddings for person identity. In: Proceedings of ECCV (2018)
27. Nagrani, A., Albanie, S., Zisserman, A.: Seeing voices and hearing faces: cross-modal biometric matching. In: IEEE CVPR (2018)
28. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
29. Nagrani, A., Yang, S., Arnab, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. In: NIPS (2021)
30. Nawaz, S., et al.: Cross-modal speaker verification and recognition: a multilingual perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1682–1691 (2021)
31. Okabe, K., Koshinaka, T., Shinoda, K.: Attentive statistics pooling for deep speaker embedding. In: Proceedings of Interspeech, pp. 2252–2256 (2018)
32. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the BMVC, pp. 41.1–41.12 (2015)
33. Praveen, R.G., Cardinal, P., Granger, E.: Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention. IEEE Trans. Biomet., Behav. Identity Sci. **5**(3), 360–373 (2023)
34. Praveen, R.G., Granger, E., Cardinal, P.: Cross attentional audio-visual fusion for dimensional emotion recognition. In: IEEE FG, pp. 1–8 (2021)
35. Praveen, R.G., Granger, E., Cardinal, P.: Recursive joint attention for audio-visual fusion in regression based emotion recognition. In: IEEE ICASSP, pp. 1–5 (2023)
36. Praveen, R.G., et al.: A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2485–2494 (2022)
37. Sarı, L., Singh, K., Zhou, J., Torresani, L., Singhal, N., Saraf, Y.: A multi-view approach to audio-visual speaker verification. In: IEEE ICASSP, pp. 6194–6198 (2021)
38. Shon, S., Oh, T.H., Glass, J.: Noise-tolerant audio-visual online person verification using an attention-based neural network fusion. In: IEEE ICASSP, pp. 3995–3999 (2019)
39. Snyder, D., Chen, G., Povey, D.: MUSAN: A music, speech, and noise corpus. CoRR abs/1510.08484 (2015). https://arxiv.org/abs/1510.08484
40. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: Proceedings of Interspeech, pp. 999–1003 (2017)

41. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust DNN embeddings for speaker recognition. In: IEEE ICASSP, pp. 5329–5333 (2018)
42. Snyder, D., et al.: The JHU speaker recognition system for the voices 2019 challenge. In: INTERSPEECH, pp. 2468–2472 (2019)
43. Sun, P., Zhang, S., Liu, Z., Yuan, Y., Zhang, T., Zhang, H., Hu, P.: Learning audio-visual embedding for person verification in the wild (2022)
44. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: IEEE Conference on CVPR, pp. 1701–1708 (2014)
45. Tao, R., Das, R.K., Li, H.: Audio-visual speaker recognition with a cross-modal discriminative network. In: Proc. Interspeech, pp. 2242–2246 (2020)
46. Tao, R., Lee, K.A., Shi, Z., Li, H.: Speaker recognition with two-step multi-modal deep cleansing. In: IEEE ICASSP, pp. 1–5 (2023)
47. Villalba, J., et al.: State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST sre18. In: Interspeech, pp. 1488–1492 (2019)
48. Wan, X., Xing, T., Ji, Y., Gong, S., Liu, C.: 3D human action recognition with skeleton orientation vectors and stacked residual BI-LSTM. In: 4th IAPR ACPR, pp. 571–576 (2017)
49. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: CVPR, pp. 12692–12702 (2020)
50. Wang, Y.: Efficient audio-visual speaker recognition via deep multi-modal feature fusion. In: Proceedings of ICCIS, pp. 99–103 (2021)
51. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: CVPR (2020)
52. Wen, P., Xu, Q., Jiang, Y., Yang, Z., He, Y., Huang, Q.: Seeking the shape of sound: an adaptive framework for learning voice-face association. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16342–16351 (2021)
53. Wen, Y., Ismail, M.A., Liu, W., Raj, B., Singh, R.: Disjoint mapping network for cross-modal matching of voices and faces. In: ICLR (2019)
54. Wu, C.H., Lin, J.C., Wei, W.L.: Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA Trans. Signal Inform. Process. **3**, e12 (2014)
55. Wu, Z., Cai, L., Meng, H.: Multi-level fusion of audio and visual features for speaker identification. In: Advances in Biometrics, pp. 493–499 (2005)
56. Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E.A., Luo, J.: Deep multimodal representation learning from temporal data. In: CVPR (2017)
57. Zeng, Z., et al.: Audio-visual affect recognition through multi-stream fused hmm for HCI. In: CVPR, pp. 967–972 (2005)
58. Zhao, X., Lv, Y., Huang, Z.: Multimodal fusion-based swin transformer for facial recognition micro-expression recognition. In: IEEE ICMA, pp. 780–785 (2022)
59. Zhao, Z., Li, Z., Wang, W., Zhang, P.: PCF: ECAPA-TDNN with progressive channel fusion for speaker verification. In: IEEE ICASSP, pp. 1–5 (2023)
60. Zhao, Z., et al.: A lighten CNN-LSTN model for speaker verification on embedded devices. Futur. Gener. Comput. Syst. **100**, 751–758 (2019)