LEARNING AUDIO-VISUAL EMBEDDING FOR WILD PERSON VERIFICATION

Peiwen Sun\textsuperscript{1,2}, Shanshan Zhang\textsuperscript{2}, Zishan Liu\textsuperscript{1}, Yougen Yuan\textsuperscript{2}, Taotao Zhang\textsuperscript{2}, Honggang Zhang\textsuperscript{1}, Pengfei Hu\textsuperscript{2}

\textsuperscript{1}School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China \quad \textsuperscript{2}TEG AI, Tencent Inc, Beijing, China

ABSTRACT

It has already been observed that audio-visual embedding can be extracted from these two modalities to gain robustness for person verification. However, the aggregator that used to generate a single utterance representation from each frame does not seem to be well explored. In this article, we proposed an audio-visual network that considers aggregator from a fusion perspective. We introduced improved attentive statistics pooling for the first time in face verification. Then we find that strong correlation exists between modalities during pooling, so joint attentive pooling is proposed which contains cycle consistency to learn the implicit inter-frame weight. Finally, fuse the modality with a gated attention mechanism. All the proposed models are trained on the VoxCeleb2 dev dataset and the best system obtains 0.18\%, 0.27\%, and 0.49\% EER on three official trial lists of VoxCeleb1 respectively, which is to our knowledge the best-published results for person verification. As an analysis, visualization maps are generated to explain how this system interact between modalities.

Index Terms— person verification, audio-visual, cycle consistency

1. INTRODUCTION

Voice and face are two typical biometric characteristics that have been investigated a lot for the person verification task. Accordingly, the speaker verification and face verification tasks are two separate research topics.

For speaker verification, research\textsuperscript{[1]} proposed attentive statistics pooling to focus on important frames of speaker verification and get higher discriminative power than the traditional averaging method. In further research, ECAPA-TDNN\textsuperscript{[2]} was based on blocks of TDNNs and Squeeze-Excitation(SE)\textsuperscript{[3]} to reconstruct frame-level features. The loss function was also customized to adapt to the speaker verification task\textsuperscript{[4,5]}. Recent advances driven by large-scale pre-trained models have taken the task of speaker recognition to a new level. With pre-training, Superb\textsuperscript{[6]}, Unispeech\textsuperscript{[7]}, Wavlm\textsuperscript{[8]}, HuBERT\textsuperscript{[9]} have achieved excellent performances on multiple sets.

For face verification, the margin-based softmax loss function is widely used for training face recognition models\textsuperscript{[10–13]}. Without the margin, learned features are not sufficiently discriminative. SphereFace\textsuperscript{[12]}, CosFace,\textsuperscript{[13]} and ArcFace\textsuperscript{[10]} introduced different forms of margin functions. Most of the research on audio-visual works\textsuperscript{[14,15]} still focused on extracting frame-level embeddings and then simply averaging them to get segment-level embeddings.

The performance of speaker or face verification systems would degrade dramatically under noise circumstances. Audio-visual identity verification came to the stage to gain more robustness. Researchers have found that simply fusing the scores from speaker embeddings and facial embeddings can yield good results\textsuperscript{[16,17]}. For the first time, S. Shon\textsuperscript{[18]} attempted to fuse audio-visual information with deep learning-based models to achieve better results. As transformer has been proposed, transformer seemed more efficient\textsuperscript{[15]} to fuse different modality in large dataset. But correlation had not been properly discussed before, Y. Liang\textsuperscript{[19]} explored the HGR maximal correlation of person verification. In the recent boom of pretrain models, B. Shi\textsuperscript{[20]} provided an audio-visual identity recognition method based on unsupervised pre-training models.

The aggregator that used to generate a single utterance representation from each frame does not seem to be well explored. The previous aggregator used either adaptive weights only in unimodal or simple averaging in audio-visual task. To change the current situation, overall network is proposed and brings a huge leap forward in this article. Here we improve the attention statistics pooling\textsuperscript{[1]} proposed in speaker verification and apply it to face verification. It achieves the best face verification performance on the VoxCeleb1\textsuperscript{[21]}. Intuitively, facial muscles and syllables of pronunciation have an impact on this task. Cycle consistency was introduced into the study for the inference of the keyframes of two modalities. The method that combines cycle consistency and attentive statistics pooling is called joint attentive polling. And the best system obtains the best published results for person verification to our knowledge.

2. METHOD

The overall network consists of backbone of each modality (Sec\textsuperscript{[3]}), joint attentive pooling module and fusion module.
(Sec 2.3). Joint attentive pooling module is the combination of improved attentive statistical pooling in Sec 2.1 and cycle consistency learning Sec 2.2.

2.1 Improved attentive statistical pooling

The data of each modality are pre-processed as stated in Sec 3 and sent to the respective encoders. But when the audio information is added for joint training, there is a serious drop in visual. To solve the problem, we introduce tanh activation. It plays a positive role in many aspects, which can be proved in visual. To solve the problem, we introduce tanh activation.

Through the original attentive statistical pooling mechanism in [2], each frame will be given different weights.

\[ e^{t,c} = (v^c)^T \tanh (W h^t + b) + k^c, \]

where \( h^t \) denotes the activations of the last frame layer at time step \( t \). The parameters \( W \in \mathbb{R}^{C \times T} \) and \( b \in \mathbb{R}^{C \times 1} \) project the information for self-attention into a smaller dimensional representation that is shared across all \( C \) channels to reduce the parameter count and risk of overfitting.

This information is transformed to a channel-dependent self-attention score through a linear layer with weights \( v^c \in \mathbb{R}^{T \times 1} \) and bias \( k^c \).

It is found that the value of attention \( \lambda_t \) is comparatively small standard deviation when dimensions except the temporal dimension are eliminated. That is to say, the keyframes in the time domain are more concentrated.

\[ \lambda_t = \text{sum}_{t} (e^{t,c}) \]

\[ \lambda_{tanh} = mu(\lambda_t) \ast \tanh \left( \frac{\lambda_t - mu(\lambda_t)}{std(\lambda_t)} \right) + mu(\lambda_t), \]

where \( mu(\cdot), std(\cdot) \) is mean and standard deviation of the matrix, \( \lambda_t \in \mathbb{R}^{T \times 1} \) denotes the temporal attention and \( e^{t,c} \in \mathbb{R}^{T \times T} \) denotes the attention map that obtain from last step. The activated weights are projected into a more dispersed and polarized space \( \mathbb{R}^{T \times 1} \) and preserve the mean. If not, fine-grained granularity brings the difficulties of convergence for further experiment. The activation method can avoid non-convergence, alleviate overfitting and be beneficial to the later learning of audio-visual weight interactions.

Then \( e^{t,c} \) attention map can be updated as

\[ e_{tanh}^{t,c} = e^{t,c} + \frac{\lambda_t}{\lambda^{t}}, \]

where \( e_{tanh}^{t,c} \in \mathbb{R}^{T \times A} \) denotes the attention map projected from \( e^{t,c} \). This scalar score \( e_{tanh}^{t,c} \) denotes then normalized over all frames by applying the softmax function channel-wise across time:

\[ \alpha_{t,c} = \frac{\exp (e_{tanh}^{t,c})}{\sum_t \exp (e_{tanh}^{t,c})}, \]

The self-attention score \( \alpha_{t,c} \) represents the importance of each frame given the channel and is used to calculate the weighted statistics of channel \( c \). Then weighted mean vector and weighted standard deviation vector can be calculated as

\[ \tilde{\mu}^c = \sum_t \alpha_{t,c} h^{t,c}, \]

\[ \tilde{\sigma}^c = \sqrt{\sum_t \alpha_{t,c} (h^{t,c} - \tilde{\mu}^c)^2}. \]

2.2 Cycle consistency learning

From the perspective of face verification, researchers such as [22] hope to eliminate the impact of facial expressions on face verification. Then, from the perspective of speaker verification, in the interval of speech, some speech factors are not helpful for speaker verification [11]. Then there is a natural idea, hoping to establish some potential probability relationships of keyframes between the face verification and speaker verification, which is called the “complementary rule”.

Essentially, we want to build a network where keyframes between different modalities can be derived from each other. And naturally, there comes a loop. The network used to construct the supervised loop is two encoders consisting of 3-layer mlp to encode the weights of simple temporal frames.

Fig. 1. Overall Network
Here we generalize adversarial loss and cycle consistency loss into regression tasks. Adversarial loss follows the previous paper [23] to serve a similar purpose, but it is not really adversarial. The adversarial loss ensures that a single encoder can evolve smoothly, which in turn ensures that the encoder can produce more realistic temporal weights. The cycle consistency loss ensures that temporal weights from different sources in the same modality have the same intensity.

Adversarial loss and cycle consistency loss only involve the mapping of temporal attention $\lambda^t_{\text{tanh}}$, and we express the objective as:

$$L_{\text{adv}} = 2 - \langle a, g(v) \rangle - \langle v, f(a) \rangle,$$

$$L_{\text{cycle}} = 2 - \langle v, f(g(v)) \rangle - \langle a, g(f(a)) \rangle,$$

while $g(\cdot)$ and $f(\cdot)$ represent 2 mlp encoders separately, $L_{\text{adv}}$ and $L_{\text{cycle}}$ denote adversarial loss and cycle consistency loss, $\langle \cdot, \cdot \rangle$ is the cosine similarity operator.

Each direction of the loop generates 2 loss which forces the attentive statistics pooling to learn potential information. Therefore, the overall loss can be calculated as

$$L = L_{\text{AAM}} + \beta L_{\text{adv}} + \gamma L_{\text{cycle}},$$

while $\beta$ and $\gamma$ are constants that are used to control and adjust the weights of different losses.

Compared with the L2 loss used by others, the supervision of cosine similarity is more weakened. Cosine similarity is more concerned with the relative relationship between the weights of different time frames of the current modality, that is, more concerned with the direction of temporal attention as a vector.

### 2.3. Fusion strategy

There have been many explorations in fusion methods in recent years. Bilinear pooling [24], transformers [25] are used to extract mutual features that have time-series input relationships or logical correspondences. In terms of identity recognition, there is no necessary relationship between the utterance level features of the face and speaker after pooling. Simple attention gate method Fig.3 determined by attention shows stronger convergence ability.

For the joint embedding output by the network, the convergence can be accelerated by some tricks during the training process. We follow [26] to impose an orthogonality constraint on the fused embeddings.

### 3. EXPERIMENT SETUP

VoxCeleb1&2 [21] [27] are used in our experiment. VoxCeleb is an audio-visual dataset consisting of 2,000+ hours short clips of human speech, extracted from interview videos on YouTube. For model training, we use the development set of VoxCeleb2, which contains 5,994 speakers. To better illustrate the advantages of our network, we adopt 3 trials in VoxCeleb1. For audio data, 80-dimensional Fbank features are extracted with a 25 ms window and 10 ms frameshift, and augmentation with the random mask is added along both the time and frequency domain. Then we do the cepstral mean on the Fbank features. The MUSAN [28] and RIR Noise datasets [29] are used as noise sources and room impulse response functions, respectively. For each video segment, we extracted 6 fps in VoxCeleb1 [21] and 25 fps in VoxCeleb2 [27] datasets. Then use the similarity transformation to map the face region to the same shape ($1 \times 128 \times 128$), which means that we use grey images instead of RGB. Finally, we normalize each image’s pixel value to reside in the range of [-0.5, 0.5].

For audio encoder, we use ECAPA-TDNN [2] into which Fbank feature is fed to extract speaker embedding. We made only a few changes to ECAPA-TDNN, which adds additional scale information to the second layer aspect. For visual encoder, Face feature is extracted by IResNet18 backbone same as [10].

The training process is divided into two stages. Two different modalities are trained separately and jointly train the fusion module and cycle for finetuning. All models are trained using AAMsoftmax loss with a margin of 0.5 and a scaling factor of 30 with 32 NVIDIA Tesla P40s. We use...
the SGD optimizer with an initial learning rate of 0.01 and decrease the learning rate by 50% every 2 epochs. \( \beta \) and \( \gamma \) are set to 1 and 0.5. We also set the weight decay as 5e-4 to avoid overfitting. The global batch size is 320.

We use cosine distance with adaptive s-norm [34] for scoring. Then we report the Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) with \( P_{\text{target}} = 0.01 \) and \( C_{FA} = C_{Miss} = 1 \) for performance evaluation.

4. RESULT ANALYSIS

From table 1, we conclude that the tanh activation for unimodality will not bring obvious changes to speech while using activation in the face alone can be a relief for the generalization pressure of attentive statistic pooling. The reason for the phenomenon comes from the fact that faces have fewer frames and fewer variations. And the later experiments show that the activation operation on a single modality brings better performance on audio-visual jointly. In terms of vision alone, compared with the results of works [14, 15], our proposed method uses less image information (grey image instead of RGB), less data preparation (no face detection and face alignment), shallower extractor (iResNet18 instead of ResNet34 [15] or SE-ResNet50 [13]). This method enable us to gain a competitive improvement of 37% compared to others by adopting higher sampling rate.

The cycle-consistency loss will finally converge to around 0.03-0.05. As the loss converges and each encoder is analyzed separately, the transformation from vision to audio is easier to converge than the transformation from audio to vision. This shows that the “complementary rule” is mainly reflected in “vision supervision audio”. From the results in Table 1, the use of cycle loss indeed enhances the weight of important time frames. Compared to audio-visual person verification, we have a huge improvement of 30% to 50% compared to the original SOTA, which provides a new baseline in audio-visual person verification.

Attention map Fig. 4 of the output of the two different layers is generated to verify the “complementary rule”. We can intuitively see that the time frame weight is adjusted

![Fig. 4](image-url). Attention Heatmap of a 0.8 sec video clip

![Fig. 5](image-url). Sample visualization

| Type          | Architecture       | #Modality | VoxCeleb1-O      | VoxCeleb1-E     | VoxCeleb1-H     |
|---------------|--------------------|-----------|------------------|------------------|------------------|
|               |                    |           | EER(%) | MinDCF | EER(%) | MinDCF | EER(%) | MinDCF |
| Unimodal      | ECAPA-TDNN [2]     | A         | 0.87   | 0.107  | 1.12   | 0.132  | 2.12   | 0.210  |
|               | SimAM-Resnet [35]  | A         | 0.64   | 0.067  | 0.84   | 0.089  | 1.49   | 0.146  |
| Z. Chen [14]  | \( \cap \)V       | 2.308    | -      | 2.234  | -      | 3.782  | -      | -      |
|               | \( \cap \)fusion   | 2.260    | -      | 1.542  | -      | 2.374  | -      | -      |
|               | ensemble(\( \cap \),\( \cup \)) | 0.505  | -      | 0.432  | -      | 0.782  | -      | -      |
|               | ensemble(\( \cap \),\( \cup \),\( \cap \)) | 0.499  | -      | 0.379  | -      | 0.683  | -      | -      |
| Audio-visual  | Y. Qian [15]       | A         | 1.62   | -      | 1.75   | -      | 3.16   | -      |
|               | V                  | 3.04     | -      | 2.18   | -      | 4.23   | -      | -      |
|               | fusion             | 0.71     | -      | 0.48   | -      | 0.85   | -      | -      |
| Ours          | A w/o tanh†        | 0.98     | 0.140  | 1.24   | 0.163  | 2.30   | 0.264  |
|               | \( \cap \)A        | 0.99     | 0.140  | 1.24   | 0.163  | 2.27   | 0.264  |
|               | V w/o tanh†        | 2.59     | 0.214  | 1.88   | 0.198  | 3.39   | 0.297  |
|               | \( \cap \)V        | 1.44     | 0.147  | 1.28   | 0.157  | 2.14   | 0.230  |
|               | fusion w/o cons†   | 0.22     | 0.022  | 0.27   | 0.035  | 0.52   | 0.058  |
|               | \( \cap \)fusion   | 0.18     | 0.017  | 0.26   | 0.035  | 0.49   | 0.057  |
| Audio-visual  | Ours               | ensemble(\( \cap \),\( \cup \),\( \cap \)) | 0.14 | 0.012 | 0.21 | 0.028 | 0.37 | 0.046 |

† “w/o cons” means “without cycle consistency” and “w/o tanh” means “without activation based on tanh”.

Table 1. Performance of proposed network
adaptively and the two modalities are roughly complementary. The most accurate time frame for face verification is often accompanied by rare facial muscle movements; the most accurate time frame for speaker verification is often accompanied by a wide range of facial muscle movements.

About 20 hard samples were selected from the VoxCeleb1-1H dataset for t-SNE dimensionality reduction visualization analysis. It can be seen from Fig.5 that each diagram of a single modality can easily find 5-6 outliers. But our network is more robust to these outliers.

Generally, we propose a novel backbone for audio-visual person verification. It significantly outperforms the popular systems in recognition performance. And analysis shows the robustness of our network and the existence of “complementary rules”. In future work, the audio-visual unsupervised pre-train model will be explored to reach a higher level.

5. REFERENCES

[1] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” arXiv preprint arXiv:1803.10963, 2018.

[2] B. Desplanques, J. Thienpondt, and K. Demunyck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in Interspeech, pp. 3830–3834, 2020.

[3] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. CVPR, pp. 7132–7141, 2018.

[4] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” in Proc. ICASSP, pp. 4879–4883, IEEE, 2018.

[5] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, “Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition,” in APSIPA ASC, pp. 1652–1656, IEEE, 2019.

[6] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-J. Lai, K. Lakhota, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, et al., “Superb: Speech processing universal performance benchmark,” arXiv preprint arXiv:2105.01051, 2021.

[7] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, et al., “Unispeech-sat: Universal speech representation learning with speaker aware pre-training,” in Proc. ICASSP, pp. 6152–6156, IEEE, 2022.

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” arXiv preprint arXiv:2110.13900, 2021.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE Trans Audio Speech Lang Process, vol. 29, pp. 3451–3460, 2021.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in Proc. CVPR, June 2019.

[11] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, “Curricularface: adaptive curriculum learning loss for deep face recognition,” in Proc. CVPR, pp. 5901–5910, 2020.

[12] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Spherface: Deep hypersphere embedding for face recognition,” in Proc. CVPR, pp. 212–220, 2017.

[13] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in Proc. CVPR, pp. 5265–5274, 2018.

[14] Z. Chen, S. Wang, and Y. Qian, “Multi-modality matters: A performance leap on voxceleb,” in INTERSPEECH, pp. 2252–2256, 2020.

[15] Y. Qian, Z. Chen, and S. Wang, “Audio-visual deep neural network for robust person verification,” IEEE Trans Audio Speech Lang Process, vol. 29, pp. 1079–1092, 2021.

[16] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, J. Hernandez-Cordero, et al., “The 2019 nist audio-visual speaker recognition evaluation...” in Odyssey, pp. 259–265, 2020.

[17] J. Alam, G. Boulianne, L. Burget, M. Dahmane, M. D. Sanchez, A. Lozano-Diez, O. Glembek, P.-L. St-Chars, M. Lalonde, P. Matejka, et al., “Analysis of acl submission to nist sre 2019 cmm and vast challenge...” in Odyssey, pp. 289–295, 2020.

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

[19] Y. Liang, F. Ma, Y. Li, and S.-L. Huang, “Person recognition with hgr maximal correlation on multimodal data,” in Proc. ICPR, pp. 1–8, IEEE, 2021.

[20] B. Shi, A. Mohamed, and W.-N. Hsu, “Learning lip-based audio-visual speaker embeddings with av-hubert,” arXiv preprint arXiv:2205.07180, 2022.

[21] A. Nagarjuna, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” Computer Science and Language, 2019.

[22] J. Chang, Z. Lan, C. Cheng, and Y. Wei, “Data uncertainty learning in face recognition,” in Proc. CVPR, pp. 5710–5719, 2020.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. CVPR, pp. 2223–2232, 2017.

[24] J.-H. Kim, J. Jun, and B.-T. Zhang, “Bilinear attention networks,” Proc. NeurIPS, vol. 31, 2018.

[25] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” in Proc. ACL, vol. 2019, p. 6558, NIH Public Access, 2019.

[26] M. S. Saeed, M. H. Khan, S. Nawaz, M. H. Yousaf, and A. Del Bue, “Fusion and orthogonal projection for improved face-voice association,” in Proc. ICASSP, pp. 7057–7061, 2022.

[27] J. S. Chung, A. Nagarjuna, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in INTERSPEECH, 2018.

[28] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” arXiv preprint arXiv:1510.08484, 2015.

[29] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in Proc. ICASSP, pp. 5220–5224, IEEE, 2017.

[30] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in Proc. CVPR, 2020.

[31] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, “Sample and computation redistribution for efficient face detection,” in Proc. ICLR, 2021.

[32] S. Yang, P. Luo, C. C. Loy, and X. Tang, “Wider face: A face detection benchmark,” in Proc. CVPR, 2016.

[33] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, “Face alignment across large poses: A 3d solution,” in Proc. CVPR, June 2016.

[34] P. Matejka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernocký, “Analysis of score normalization in multilingual speaker recognition...” in Interspeech, pp. 1567–1571, 2017.

[35] X. Qin, N. Li, C. Weng, D. Su, and M. Li, “Simple attention module for deep speaker embedding,” arXiv preprint arXiv:2110.13900, 2021.