CAFNet: Cross Attentive Face-Voice Fusion Network

Lakshya Gupta
Dept. of Mechanical Engineering
IIT Kanpur
lakshyagarg911@gmail.com

Pratham Gupta
Dept. of Electrical Engineering
IIT Kanpur
prathamgupta1232@gmail.com

Shubham Yadav
Dept. of Mechanical Engineering
IIT Kanpur
shubhyd13@gmail.com

***Abstract*—We present CAFNet, a cross-attentive fusion network that aligns and fuses speaker and face representations for robust face-voice verification in multilingual settings. Our pipeline combines audio preprocessing with strong modality-specific embeddings and a bidirectional cross-attention fusion head trained with cross-entropy. We report progress-phase results on the English–Urdu dataset of MAV-Celeb: an initial WavLM-based pipeline yielded an overall EER $\approx$ 0.2702, while the final ECAPA-TDNN + CAFNet pipeline improved this to 0.26316. The proposed architecture consistently improves cross-modal alignment and is compact enough to train on limited data.**

Code available at: https://github.com/prathamg007/Face_Voice_Association

***Index Terms*—face-voice association, cross-attention, ECAPA-TDNN, VGGFace2, diarization, multimodal fusion**

I. Introduction

Face-voice association seeks to verify whether a face image and a speech segment belong to the same identity. Real-world audio-visual data are noisy, multilingual and frequently multi-speaker; robustness to these conditions is critical for deployed systems. We design CAFNet to explicitly model cross-modal interactions via bidirectional cross-attention, and to combine this with a practical audio preprocessing pipeline to improve representation quality. Our aims are (1) to obtain a discriminative fused representation robust to language shifts and noise, and (2) to remain consistent with best practices for training on modest-sized multilingual datasets.

II. Related Work

Prior work on face-voice matching typically follows a two-branch paradigm: modality-specific encoders produce embeddings which are then fused via learned projections or simple concatenation [13], [14]. The FOP family of methods introduced orthogonality-promoting losses and gated fusion to improve separation across identities [1], [10]. Recent trends favor stronger backbone encoders for each modality and attention-based fusion to explicitly model cross-modal dependencies [11], [12]. Our approach builds on these ideas by adding a bidirectional cross-attention block (face queries voice and voice queries face), combined with a gated fusion stage and supervised cross-entropy training.

III. Dataset and Preprocessing

Dataset. We trained on the English-Urdu MAV-Celeb dataset. Baseline values for comparison were taken from the official FAME baseline FOP paper [9].

Audio preprocessing. We singled out the primary speaker from the conversation audios, we performed speaker diarization, and assumed the speaker with maximum conversation time as the primary speaker. In this process, all the audio files within a subgroup for a particular speaker are concatenated to form a single, continuous stream. We then utilized the pre-trained pyannote.audio [7] [8] v3.1 speaker diarization pipeline. The primary speaker for each subgroup was then identified, as the one with the maximum total speech duration, as given by the pyannote.audio model. Finally, all the speech segments belonging to the primary speaker were then extracted, concatenated, and divided back into the same number of files as in pre diarization process.

IV. Methodology

A. Backbone embeddings

We extract fixed-length modality embeddings:

- **Face:** VGGFace2 embeddings (per-image pooled features) [5].
- **Audio:** speaker embeddings extracted using WavLM (early experiments) [4] and ECAPA-TDNN (final pipeline) [3].

All embeddings are L2-normalized before projection.

B. Projection and cross-attention

Both audio and face vectors are projected to a shared latent dimension d via learned linear layers:

$$z_a = \tanh(W_a x_a), \quad z_v = \tanh(W_v x_v), \quad z_a, z_v \in \mathbb{R}^d.$$

We add a sequence dimension of length 1 and apply bidirectional multi-head attention:

$$\text{Attn}_{a\to v} = \text{MHA}(Q = z_a,\ K = z_v,\ V = z_v),$$

$$\text{Attn}_{v\to a} = \text{MHA}(Q = z_v,\ K = z_a,\ V = z_a).$$

Residual connections and layer normalization follow each cross-attention output:

$$\tilde{z}_a = \text{LN}(z_a + \text{Attn}_{v\to a}), \quad \tilde{z}_v = \text{LN}(z_v + \text{Attn}_{a\to v}).$$

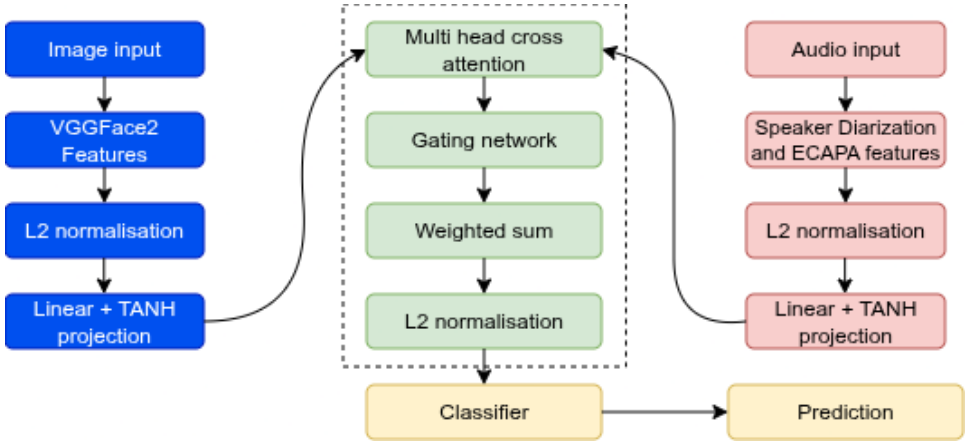

Fig. 1. The proposed CAFNet architecture illustrating the bidirectional cross-attention mechanism between face and audio embeddings.

TABLE I
PERFORMANCE COMPARISON ON THE MAV-CELEB (ENGLISH URDU) DEVELOPMENT SET

Method	Overall EER
FAME Baseline (FOP)	33.4
WavLM + CAFNet (progress)	27.02
ECAPA-TDNN + CAFNet (final, progress)	26.316

TABLE II
PERFORMANCE COMPARISON ON THE MAV-CELEB (ENGLISH GERMAN) DATASET

Method	Overall EER
Baseline (FOP)	40.2
ECAPA-TDNN + CAFNet (final)	39.46

Note: Results indicate EER (in %), with lower values indicating a better performance.

C. Gated fusion and classifier

A learned sigmoid gate $\gamma = \sigma(W_g[\tilde{z}_a \| \tilde{z}_v])$ computes relative weighting, and the final fused vector is:

$$z_f = \gamma \odot \tilde{z}_a + (1 - \gamma) \odot \tilde{z}_v.$$

A compact MLP projects z_f to logits for binary same/different classification. We train with standard cross-entropy:

$$\mathcal{L}_{CE} = -\sum_i y_i \log(\hat{p}_i).$$

V. IMPLEMENTATION DETAILS

- **Projection dim:** $d = 256$ (experimented with 256 and 512).
- **Attention:** 8 heads, batch-first MultiheadAttention (PyTorch).
- **Training:** AdamW, LR $= 2 \times 10^{-4}$ (head), weight decay $1e{-}4$, early stopping on validation EER.
- **Sampling:** per-identity pair sampling with negative ratio 1.5; data augmentation for audio (noise, gain, reverberation) is applied during training.
- **Evaluation:** verification EER computed from soft score (probability of same identity).

VI. EXPERIMENTS

Progress-phase runs. We ran two main pipelines on the English–Urdu development split:

- **WavLM backbone + CAFNet (preprocessing applied):** overall EER $\approx$ **27.02**.
- **ECAPA-TDNN backbone + CAFNet (final):** overall EER $\approx$ **26.316**.

We also tried experimenting with processing audio with various filters, but most of them ended up removing some kind of information, leading to loss in accuracy.

A. Ablations and observations

- **Preprocessing:** diarization, to help with cleaning of the audio data, remove no speech sections, and make sure that only the part where the main speaker speaks is fed into the model.
- **Fusion:** cross-attention consistently outperformed simple gated fusion in our trials, especially under language mismatch (heard $\rightarrow$ unheard).
- **Overfitting risk:** cross-attention increases model capacity — strong regularization (dropout, augmentation, weight decay) and freezing backbones early are recommended.

VII. CONCLUSION

We introduced CAFNet, a compact cross-attention fusion head for face–voice verification with a practical audio preprocessing pipeline. CAFNet improves cross-modal alignment and yields better EERs than a WavLM-based pipeline when using ECAPA-TDNN embeddings. Future work will add orthogonality constraints, advanced sampling (hard negatives), and language-adversarial regularization to further reduce EER under unseen-language conditions.

REFERENCES

[1] M. Saeed et al., "Fusion and Orthogonal Projection for Improved Cross-Modal Embeddings," Proc. ICASSP, 2022.

[2] D. Snyder et al., "Speech representations for speaker recognition," Interspeech, 2020.

[3] S. Desplanques, K. Wouters, and J. De Neve, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," Interspeech, 2020.

[4] O. Wang et al., "WavLM: Large-Scale Self-supervised Pre-training for Full Stack Speech Processing," ICASSP, 2022.

[5] Q. Cao et al., "VGGFace2: A dataset for recognising faces across pose and age," FG, 2018.

[6] MAV-Celeb project pages (dataset description).

[7] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," Proc. Interspeech, 2023.

[8] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," Proc. Interspeech, 2023.

[9] FAME Challenge Organizers, "FAME Challenge: Face-Voice Association in the Wild," arXiv:2508.04592, 2025.

[10] M. S. Saeed et al., "Single-branch network for multimodal training," Proc. ICASSP, 2023.

[11] M. S. Saeed et al., "A synopsis of FAME 2024 challenge: Associating faces with voices in multilingual environments," Proc. ACM Multimedia, 2024.

[12] A. Hannan et al., "PAEFF: Precise Alignment and Enhanced Gated Feature Fusion for Face-Voice Association," arXiv:2505.17002, 2025.

[13] S. Nawaz et al., "Deep latent space learning for cross-modal mapping of audio and visual signals," Proc. DICTA, 2019.

[14] S. H. Shah et al., "Speaker recognition in realistic scenario using multimodal data," Proc. ICAI, 2023.