FAME Audio-Visual Team System Description

Wuyang Chen, Yanjie Sun, Kele Xu, Yong Dou

1 METHODOLOGY

This paper proposes a contrastive learning-based chaining-cluster method for the Face-Voice Association in Multilingual Environment (FAME) 2024 challenge. Our approach begins by utilizing Supervised Cross-Contrastive (SCC) learning to establish associations between voices and faces (as detailed in subsection 1.1). Subsequently, the learned face and voice representations are employed in a post-processing phase that incorporates the chaining-cluster rescore technique, with the aim of addressing outliers prevalent in wild data (as elaborated in subsection 1.2). Considering that there are four different test scenarios in FAME (trained on Urdu and tested on Urdu; trained on Urdu and tested on English; trained on English and tested on Urdu; trained on English and tested on Urdu), for clarity, we will use one scenario (trained on Urdu and tested on Urdu) as an example.

1.1 Supervised Cross-Contrastive Learning

To enhance the understanding of the association between voice and facial features of the same identity, we propose a novel network architecture grounded in SCC learning. This architecture comprises two distinct branches, denoted as E_v and E_f, each containing 11 Transformer layers. Additionally, a single Transformer layer with shared weights is positioned between the voice and face branches. This shared-weight layer improves learning efficiency and fosters the model's ability to recognize common patterns across voice and facial data. We utilize the following loss function:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log \frac{exp(E_v(v_i)^T E_f(f_i)/\tau)}{\sum_{j=1}^{N} exp(E_v(v_i)^T E_f(f_j)/\tau)}\right] \quad (1)$$

where N is the batch size. τ is the temperature hyper-parameter. v_i and f_i represent the voice and face that belong to the same identity, while the denominator is the sum of all samples in the current batch. The trained SCC network is used to extract representations from test pairs. The initial score of the test pair (v_i, f_i) is calculated using $1 - cos(E_v(v_i), E_f(f_i))$. Where $cos(\cdot)$ is a normalized cosine similarity operation.

1.2 Chaining-Cluster

The rationale for implementing clustering is based on the fact that voice-face pairs are collected in uncontrolled environments, leading to numerous potential distractions. For audio data, these distractions may include background music, overlapping voices from other individuals, and environmental noise. For image data, distractions might involve varying facial orientations, makeup, and lighting conditions. Directly applying our SCC model to derive cross-modal similarity scores can result in inaccuracies due to these distractions. Clustering offers a robust solution by effectively managing minor distractions. For example, if an audio clip of a target male individual contains overlapping voices from females, the clustering process can still yield accurate results. This is because the primary component of the audio segment remains the male voice, with the female voices contributing minimally.

Our Chaining-cluster pipeline consists of four primary steps, as delineated in Algorithm 1. These steps will be elaborated upon subsequently in this section. For illustrative purposes, we will use the SCC model trained on Urdu and corresponding pairs from the Urdu test set as examples throughout this discussion.

Algorithm 1 Algorithm of Chaining-cluster post-processing.

Input test pairs $TEST$, initial $score$,
voice, face representations $\{V, F\}$ from SCC
Output refined $score$

for each modality m in $\{V, F\}$ **do**
  Cluster m into Male and Female clusters $S^m = \{S^m_{\mathcal{M}}, S^m_{\mathcal{F}}\}$
  for each sample i in S^m **do**
    if Distance of i to gender cluster center > T^m **then**
      $\widehat{S^m} \leftarrow S^m \setminus \{i\}$
    end if
  end for
  for cluster $S^{m,g}$ in $\{\widehat{S^m_{\mathcal{M}}}, \widehat{S^m_{\mathcal{F}}}\}$ **do**
    Cluster $S^{m,g}$ into identity clusters $S^{m,g} = \{S^{m,g}_1, ..., S^{m,g}_n\}$
    for each sample j in $S^{m,g}$ **do**
      if Distance of j to identity cluster center > T^g **then**
        $\widehat{S^{m,g}} \leftarrow S^{m,g} \setminus \{j\}$
      end if
    end for
    Compute prototype $p^{m,g}$ for clusters in $\widehat{S^{m,g}}$
  end for
end for
for each gender g in $\{\mathcal{M}, \mathcal{F}\}$ **do**
  $sim^g \leftarrow$ compute Similarity$(p^{m=V,g}, p^{m=F,g})$
  for (v, f) in $TEST$ **do**
    if $(v, f) \in S^{m,g}$ **then**
      $score_{v,f} = score_{v,f} \downarrow$ if $sim^g(v, f) > T^\alpha$
    end if
  end for
end for
for (v, f) in $TEST$ **do**
  if $(v, f) \in S^m$ **then**
    C_v = clusterCenter(v), C_f = clusterCenter(f)
    $score_{v,f} = score_{v,f} \uparrow$ if $C_v \neq C_f$
  end if
end for

Unpublished work.

In Proceedings of ACM Conference, Australia

ACM ISBN
https://doi.org/

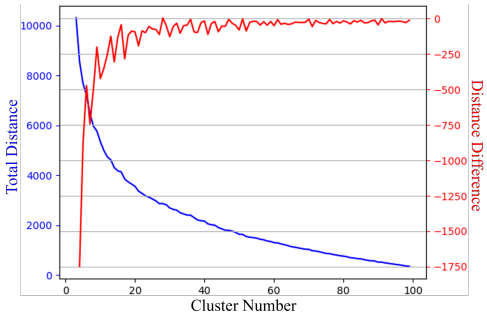

Figure 1: The blue line represents the total L_2 distance of each sample to its cluster center, and the red line represents the distance difference between adjacent cluster numbers.

Gender Cluster. For the voice-face test pairs, we initially cluster each modality into two clusters, intended to represent male ($\mathcal{M}$) and female ($\mathcal{F}$) categories. For the audio modality, we utilize the SCC model obtained from the first stage to extract audio representations $V = \{E_v(v_i)\}_{i=1}^{|\text{TEST}|}$ using its voice branch. Subsequently, we apply the K-Means algorithm to these representations, clustering them into two categories. Similarly, for the image modality, we extract face representations $F = \{E_f(f_i)\}_{i=1}^{|\text{TEST}|}$ using the SCC face branch and apply K-Means for clustering. For each modality, we then calculate the L_2 distance of each sample i to its respective cluster center. If this distance exceeds the predefined threshold T^m, we remove those samples from the current candidate set S^m, thereby forming a high-confidence set $\widehat{S^m}$.

Identity Cluster. After obtaining the gender cluster candidates for each modality, we aim further to cluster identity in each modality and gender candidate set $S^{m,g}$. For instance, consider the face modality from the female candidate set $\widehat{S^m_{\mathcal{F}}}$. We apply K-Means to test samples exist in $\widehat{S^m_{\mathcal{F}}}$ into n groups. The number of clusters n, is determined by both the Elbow method and test statistics.

The Elbow method uses the total sum of L_2 distances of each sample to its cluster center as an indicator to find the optimal number of clusters. The clustering results for $\widehat{S^m_{\mathcal{F}}}$ are shown in Figure 1. We select the optimal cluster number n by considering the index of the large difference value that is closer to half the total number of test pairs (we halve the count because we assume the numbers of male and female tests are nearly equal). In our experiment, we set $n = 16$ for heard Urdu test pairs.

Voice-Face Prototype Similarity. Next, we propose a prototype-based cross-modal similarity metric. First, the prototypes $p^{m,g}$ is calculated by averaging the representations in high confidence cluster set of $\widehat{S^{m,g}}$. The similarity matrix is then computed through the normalized cosine similarity between voice prototypes and face prototypes:

$$Sim(p^{m=V,g}, p^{m=F,g}) = \frac{E_v(p_i^V) \cdot E_f(p_j^F)}{||E_v(p_i^V)|| \cdot ||E_f(p_j^F)||} \tag{2}$$

where p_i^V and p_j^F represent the corresponding averaging voice and face representations from high confidence sets. The similarity matrix of these two modal prototypes, $sim^{g=\mathcal{F}}$ when the gender is chosen as female, is visualized as shown in Figure 2. The brighter color

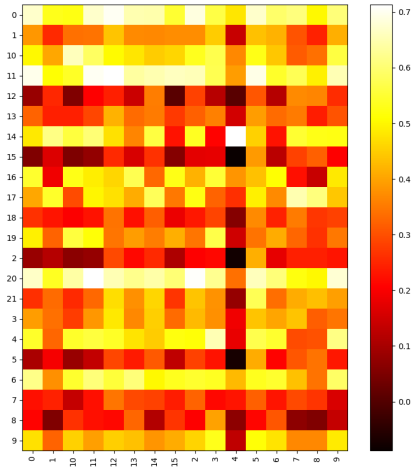

Figure 2: Cross-modal prototype similarity matrix. The raw represents voice prototypes, and the column represents the face prototypes.

indicates a higher similarity between the corresponding voice and face prototypes, suggesting a high probability that the voice-face test pairs belonging to these highly similar prototypes are from the same person. This assumption is used as a principle to refine our initial test score.

Score Refine. After previous steps, we have obtained gender clusters and identity clusters. We propose score refine method based on two principles:

- Penalize pairs of voice and face gender mismatches based on gender clusters.
- Reward pairs with high cross-modal prototype similarity based on identity clusters.

First, We compute the upper bound $B_u(s)$ and lower bound $B_l(s)$ of score base on the maximum and minimum value of initial score. Then, we perform a gender mismatch penalty:

$$score(v_i, f_i) = \begin{cases} B_u(s) & if\ C_{v_i} \neq C_{f_i} \\ score(v_i, f_i) & else \end{cases} \tag{3}$$

If a test pair (v_i, f_i) is found in the high-confidence set $\widehat{S^m}$, we determine its nearest cluster centers, denoted as C_v and C_f. A gender mismatch is identified if $C_v \neq C_f$. In such cases, the upper bound is employed to adjust the score accordingly.

Finally, we perform cross-modal prototype high similarity rewarding using the following formula:

$$score(v_i, f_i) = \begin{cases} B_l(s) + [score(v_i, f_i) - B_l(s)] * \alpha & if\ sim^g(v, f) < T^\alpha \\ socre(v_i, f_i) & else \end{cases} \tag{4}$$

For a given test pair (v_i, f_i) that exists in $S^{m,g}$, if the similarity between the corresponding voice and face prototype, denoted as $sim^g(v, f)$, is lower than the predefined threshold T^α, the distance fraction is reduced as a reward. Here, α serves as a refinement factor.