Seeking the Shape of Sound: An Adaptive Framework for Learning Voice-Face Association

Peisong Wen[1,2] Qianqian Xu[1,*] Yangbangyan Jiang[3,4]
Zhiyong Yang[3,4] Yuan He[5] Qingming Huang[1,2,6,7,*]
[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[2] School of Computer Science and Tech., University of Chinese Academy of Sciences, Beijing, China
[3] State Key Laboratory of Info. Security (SKLOIS), Inst. of Info. Engin., CAS, Beijing, China
[4] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[5] Alibaba Group, Beijing, China
[6] BDKM, University of Chinese Academy of Sciences, Beijing, China
[7] Peng Cheng Laboratory, Shenzhen, China
{wenpeisong20z,xuqianqian}@ict.ac.cn {jiangyangbangyan,yangzhiyong}@iie.ac.cn
heyuan.hy@alibaba-inc.com qmhuang@ucas.ac.cn

Abstract

*Nowadays, we have witnessed the early progress on learning the association between voice and face automatically, which brings a new wave of studies to the computer vision community. However, most of the prior arts along this line **(a)** merely adopt local information to perform modality alignment and **(b)** ignore the diversity of learning difficulty across different subjects. In this paper, we propose a novel framework to jointly address the above-mentioned issues. Targeting at **(a)**, we propose a two-level modality alignment loss where both global and local information are considered. Compared with the existing methods, we introduce a global loss into the modality alignment process. The global component of the loss is driven by the identity classification. Theoretically, we show that minimizing the loss could maximize the distance between embeddings across different identities while minimizing the distance between embeddings belonging to the same identity, in a global sense (instead of a mini-batch). Targeting at **(b)**, we propose a dynamic reweighting scheme to better explore the hard but valuable identities while filtering out the unlearnable identities. Experiments show that the proposed method outperforms the previous methods in multiple settings, including voice-face matching, verification and retrieval.*

1. Introduction

Voice and face share various potential characteristics, *e.g.*, gender, ethnicity, age, which are helpful for identification and matching. Literatures [23, 10, 15] show that humans can hear the voice of an unknown person and match the corresponding face with higher accuracy than chance, and vice versa. From the perspective of brain science, multimodal brain regions exist in the human brain, which process both voices and faces to form person identity representations [26]. Can machines learn such ability to recognize the face with the same identity only by hearing the voice, or recognize the voice from the face? In recent years, researchers have started to seek an answer to this interesting question [17, 30]. The research of this technology is beneficial to many application scenarios, including criminal investigation, synthesis or retrieval of human faces from voices [19, 31, 2, 3], *etc*. This task can be specialized as cross-modal matching, verification and retrieval problems. Different from the audio-visual speech recognition task [34], the voice-face association problem is aim to find the identity relationship between face and voice, rather than the relationship between voice and facial action.

In recent years, we have witnessed some progress of early studies along this line. As a representative example, SVHF [17] regards the matching problem as a binary classification problem, and has achieved comparable performance with human baseline in both voice-to-face matching and face-to-voice matching. Benefit from the development of deep learning and the cross-modal retrieval technology, some recent work [11, 27, 8, 33, 16] has further verified the feasibility of this problem through deep metric learning. Wen *et al*. [30] boost the performance with multiple supervision.

Despite previous methods have been able to reach the

*Corresponding authors

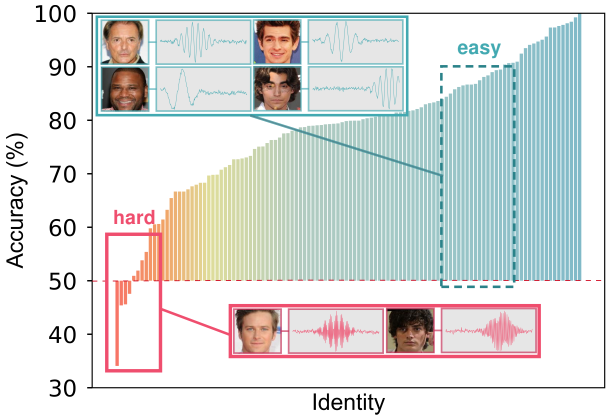


Figure 1. Accuracy of different identities in the validation set under the 1:2 voice-to-face matching setting. There is a significant gap between identities, performances of some identities are even lower than chance (50%).

same level as untrained humans, there are still two problems in learning voice-face association. **(a)** The first problem is that contrastive loss functions used in previous work only use local information in a mini-batch, which leads to slow convergence. **(b)** The second one is that the diversity of difficulty across identities is ignored. Here the diversity of difficulty means that there are obvious differences in the difficulty of learning voice-face association among different subjects. To illustrate this problem, we train a model and test its accuracy on 1:2 voice-to-face matching for different subjects. The result is shown in Fig. 1, from which it can be noticed that the accuracy of identities is significantly different. This phenomenon coincides with what we find in reality. For example, not all male voices are low and rough, and there is an unignorable fraction of male voices that have their own characteristics. In Fig. 1, we show some easy and hard identities in the obtained results. We observe that the accuracy distribution is relatively uniform, which validates our assumption that the learning difficulty is diverse. Moreover, there exists an unignorable fraction of identities suffering from a accuracy worse than random guess. This validates the existence of extremely hard identities. Identities of this kind are hardly learnable. What is worse, they might even confuse the model and shift the correct decision boundary. In this paper, we name the hard but learnable identity as *hard identity*, and the extremely hard identity as *personalized identity*. In this sense, a reasonable learning method should explore deeper into the hard identities while filtering out the personalized ones.

Based on the above consideration, in this paper, we propose an adaptive framework for the voice-face association learning. To overcome **(a)**, we introduce a two-level modality alignment, which consists of implicit and explicit modality alignment. The implicit alignment is implemented with an identity-classification-driven loss. With the theoretical analysis, we show that minimizing the implicit alignment loss could maximize the distance of embedding across modalities and identities and minimize the distance of embeddings across modalities but belong to the same identity. Moreover, the distance is measured from a global perspective instead of a local mini-batch. In this way, the implicit alignment introduces global information and identity semantics in the embeddings. Moreover, the explicit alignment, as a complementary component, aligns the two modalities in a mini-batch directly. For **(b)**, we propose an adaptive framework to handle the hard identities and personalized identities with dynamic identity weights. The hard identities obviously contribute to the bottleneck of the performance of the learning methods. We propose an adaptive weighting strategy to gradually increase the weights of the hard identities. This encourages the network to dive deeper into the associations between voice and face. Since personalized identities are extremely hard to learn, and their gradients are larger throughout the training phase, forcing the model to learn these samples will reduce the generalization of the model. Therefore, our proposed strategy adaptively assigns zero weights to personalized identities.

In a nutshell, the main contributions of this work can be summarized as follows:

- We propose explicit modality alignment and implicit modality alignment to effectively learn the voice-face association in a comprehensive manner.
- We propose an adaptive identity re-weighting framework to better explore cross-modal associations from hard identities, and excluding personalized identities for generalization.
- Experiments under various settings are conducted to illustrate the effectiveness of the proposed framework.

2. Related Work

2.1. Learning voice-face association

In recent years, learning the voice-face association has aroused the interest of researchers. To the best of our knowledge, SVHF [17] is the first work to propose a machine learning algorithm for the voice-face association learning, which focuses on the task of matching voices and faces. It poses the matching as a binary classification problem and simply uses the concatenation of the features as input of the classifier. Then the researchers turn to solve the problem from the cross-modal learning perspective, and include additional tasks such as cross-modal verification and retrieval. Inspired by existing cross-modal retrieval methods, some studies [16, 11, 33] adopt the metric learning technique, *e.g.*, contrastive loss [7] or triplet loss [29], to help bridge

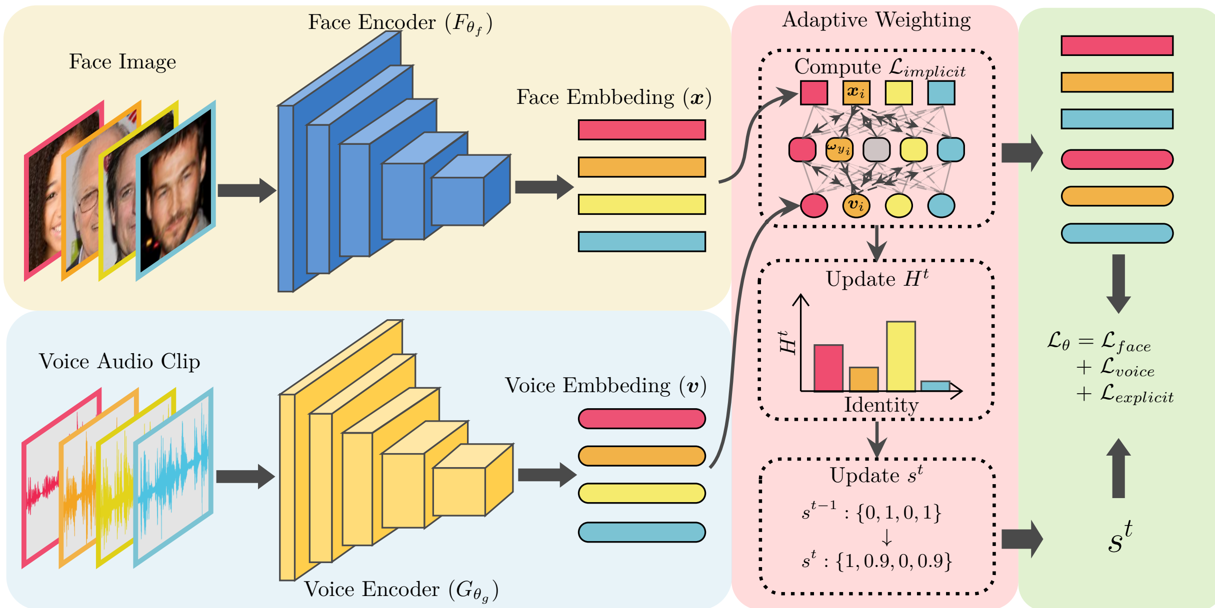


Figure 2. An overview of the proposed method. Face images and voice audio clips are fed into the face encoder network and the voice encoder network, respectively. The extracted embeddings are then assigned different weights according to the average loss of identities, and personalized samples are filtered out. Finally, the parameters of the network are updated with two-level modality alignment.

the semantic gap between the two modalities. To further exploring the relationship between samples in a mini-batch, Horiguchi *et al*. utilize N-pair loss [8, 24] and Wang *et al*. propose ranking loss [27] in the learning. Different from these approaches that are only supervised by the identity information, Wen *et al*. explicitly introduce more information like gender and nationality to supervise the learning in a multi-task manner [30].

A shortcoming of most existing metric learning methods is that the losses they adopted only utilize the local information in a mini-batch. Yet in this work, we consider a two-level modality alignment where the global information in the whole dataset and the local information in the mini-batch are simultaneously employed.

2.2. Sample re-weighting in deep learning

In deep learning, different samples have different effects on model performance. To evaluate how samples affect models, Koh & Liang [12] use influence function to test the influence of sample weighting on loss without retraining. Based on the influence function, Wang *et al*. [28] locate samples which are not helpful after the first round of training, and retrain the model from scratch without these samples. For the same purpose, Ren *et al*. [21] utilize meta learning method to assign weights based on the gradient direction of samples, and Fan *et al*. [5] explores deep reinforcement learning to automatically select samples in the training process. Sample re-weighting strategies are also applied in the object detection task [22, 14, 13]. One of the most well-known works is focal loss [14], which alleviates class imbalance with a weighted loss function.

These re-weighting strategies only consider the instance-level difficulty, thus weight each sample separately. However, in the voice-face association learning task, the difficulty is identity-level. To address this issue, we propose a novel strategy that assigns identity-specific weights, where personalized identities are further excluded with zero weights adaptively. The advantages are two-fold: on one hand, it can reduce the number of weights and thus reduce the algorithm's complexity; on the other hand, the exclusion of personalized identities ensures the generalization of the network.

3. Methodology

Our goal is to learn generic vector representations to bridge the semantic gap across voices and faces such that a series of tasks ranging from face-voice/voice-face matching, verification to retrieval could be made available. To achieve this goal, we consider two factors. First of all, to narrow the gap across modalities, we expect to perform effective modality alignment. To this end, we introduce a loss-driven alignment mechanism in Sec. 3.1 where the semantic gap is indirectly reduced by minimizing the identity classification loss function, which is called implicit modality alignment loss. Besides indirect information, we also consider an explicit modality alignment loss in Sec. 3.2 based on the contrastive learning framework. As for the second factor, we consider the diversity of learning difficulty mentioned in Sec. 1. It leads us to an adaptive learning framework with dynamic identity weights, which will be described in Sec. 3.3.

An overview of our framework is provided in Fig. 2, where we adopt the convolutional networks as our backbone. After extracting embeddings, we perform an adaptive learning procedure via identity re-weighting. The loss function is the sum of the explicit and implicit modality alignment loss. When the identity weights are obtained, the final model is obtained by retraining the network based on the weights.

3.1. Identity Recognition and Implicit Modality Alignment

First of all, we start with an implicit formulation of the modality alignment loss. To make sure that the face/voice embeddings are consistent with the semantics of identity, we expect that the learned embeddings should lead to accurate identity recognition. This motivates us to minimize the softmax loss of the identity classification for both face and voice embeddings. Recalling Fig.2, our backbone includes the deep encoders to leverage the face/voice embeddings and a linear classifier, i.e., the last FC layer of the architecture. We assume both voice and face embeddings share a common identity linear classifier. Its weight matrix is denoted as $\boldsymbol{W} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \cdots, \boldsymbol{\omega}_M] \in \mathbb{R}^{D\times M}$, where M and D represent the number of identities and feature dimensions respectively. Given the training data $\mathcal{D} = \{(\boldsymbol{v}_i, \boldsymbol{x}_i, y_i)\}_{i=1}^N$, the identity classification loss is presented as follows:

$$\mathcal{L}_{face} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\boldsymbol{\omega}_{y_i}^T \boldsymbol{x}_i)}{\sum_{j=1}^{M}\exp(\boldsymbol{\omega}_j^T \boldsymbol{x}_i)} \tag{1}$$

$$\mathcal{L}_{voice} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\boldsymbol{\omega}_{y_i}^T \boldsymbol{v}_i)}{\sum_{j=1}^{M}\exp(\boldsymbol{\omega}_j^T \boldsymbol{v}_i)} \tag{2}$$

$$\mathcal{L}_{implicit} = \mathcal{L}_{face} + \mathcal{L}_{voice} \tag{3}$$

As an interesting fact, we can prove that adopting a common classifier could lead to an implicit modality alignment mechanism. This is shown in the following proposition.

Proposition 1 *Supposing that, for any $k \in \{1, 2, \cdots, M\}$, the weight decay strategy ensures $\|\boldsymbol{\omega}_k\| \le C$, we have a lower bound of $\mathcal{L}_{implicit}$ written as follows:*

$$\mathcal{L}_{implicit} \ge 2\log M - \frac{C}{MN}\sum_{j=1}^{M} D_j$$

where

$$D_j = \left\| (M-1)\sum_{y_i=j}(\boldsymbol{x}_i + \boldsymbol{v}_i) - \sum_{y_i\neq j}(\boldsymbol{x}_i + \boldsymbol{v}_i) \right\|.$$

Prop. 1 shows the following properties of $\mathcal{L}_{implicit}$:

- According to the inequality in this proposition, minimizing $\mathcal{L}_{implicit}$ leads a smaller value of its lower bound $\log M - \frac{C}{MN}\sum_{j=1}^{M} D_j$, which equivalently leads to a larger value of $\frac{C}{MN}\sum_{j=1}^{M} D_j$.

- For a fixed j, D_j is the overall distance between the face and voice embeddings belonging to j, and the embeddings that do not belong to j. Maximizing $\frac{C}{MN}\sum_{j=1}^{M} D_j$ eventually enforces all $\boldsymbol{x}_i, \boldsymbol{v}_i$ from the same class to be close to each other and enforces all $\boldsymbol{x}_i, \boldsymbol{v}_i$ from different classes to be far away from each other. In this sense, minimizing $\mathcal{L}_{implicit}$ provides an *implicit modality alignment* mechanism for our method.

- Compared with directly maximizing $\frac{C}{MN}\sum_{j=1}^{M} D_j$, minimizing $\mathcal{L}_{implicit}$ introduces $\boldsymbol{W}$ to the model which leverages global information shared across all sample points. From the efficiency perspective, optimizing D_k directly requires traversing the entire training set, while the implicit optimization can be accelerated by mini-batch training.

3.2. Explicit modality alignment

In the previous subsection, we show that the identity recognition loss behaves like a global and implicit modality alignment loss. To obtain a comprehensive loss, we introduce a local and explicit modality alignment loss to include the complementary information of the implicit loss.

Instead of using paired data directly, we implement the explicit modality alignment with N-pair loss [24, 25], which explores the local relationship among a mini-batch of instances. It is formulated as follows:

$$\begin{aligned}\mathcal{L}_{explicit} = &\frac{1}{N}\sum_{i=1}^{N}\log\Big(m + \frac{\sum_{y_j\neq y_i}\exp(\boldsymbol{v}_i\hat{\boldsymbol{x}}_j)}{\exp(\boldsymbol{v}_i\hat{\boldsymbol{x}}_i)}\Big) \\ &+ \frac{1}{N}\sum_{i=1}^{N}\log\Big(m + \frac{\sum_{y_j\neq y_i}\exp(\boldsymbol{x}_i\hat{\boldsymbol{v}}_j)}{\exp(\boldsymbol{x}_i\hat{\boldsymbol{v}}_i)}\Big)\end{aligned} \tag{4}$$

where m is a hyper parameter to control inter-class margins, and $\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$, $\hat{\boldsymbol{v}} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$.

This loss function adds explicit constraints on the embeddings of the two modalities, as a powerful complementary to the implicit alignment. At first glance, the ratio of positive samples and negative samples is $1 : (\mathrm{N}-1)$, which might lead to the imbalance issue. Nonetheless, according to the following analysis, we can see that the loss terms will not be affected by the overwhelming ratio of the negative instances. In fact, $\mathcal{L}_{explicit}$ can be approximately written as

$$\begin{aligned}\mathcal{L}_{explicit} \approx \frac{1}{N}\sum_{i=1}^{N}[\max_{y_j \neq y_i}\{\boldsymbol{v}_i\hat{\boldsymbol{x}}_j\} - \boldsymbol{v}_i\hat{\boldsymbol{x}}_i + m - 1]_+ \\ + \frac{1}{N}\sum_{i=1}^{N}[\max_{y_j \neq y_i}\{\boldsymbol{x}_i\hat{\boldsymbol{v}}_j\} - \boldsymbol{x}_i\hat{\boldsymbol{v}}_i + m - 1]_+\end{aligned} \tag{5}$$

where $[x]_+$ indicates $\max(x, 0)$. This shows that $\mathcal{L}_{explicit}$ behaves like a hinge loss to punish $\max_{y_j \neq y_i}\{\boldsymbol{v}_i\hat{\boldsymbol{x}}_j\} - \boldsymbol{v}_i\hat{\boldsymbol{x}}_i$. In this sense, only the one negative instance realizing the maximum is activated for each loss term. Hence, the imbalance issue could be naturally avoided.

3.3. Adaptive identity re-weighting

In this subsection, we propose an adaptive learning algorithm to deal with the diversity of the learning difficulties. The proposed method has three stages. During the first stage, we perform a warm-up training, where a pretrained model is learned without identity weights. In the second stage, the pretrained model is updated together with the identity weights. In the third stage, we train the final model based on the learned identity weights. A summary of all the details is shown in Alg. 1.

Algorithm 1 Training with identity weights

Input: Training data $\mathcal{D}$, warm up iteration T_{warm}, update iteration T_{update}, max iteration T_{max}, batch size N, number of identities M, ratio of data retained R_{keep}.
Output: model parameters θ_f, θ_g.

1: ▷ `First Stage`
2: `Train` $F_{\theta_f}, G_{\theta_v}$ `with` $\mathcal{D}$ `for` T_{warm} `iterations.`
3: ▷ `Second Stage`
4: **while** $\sum_{i=0}^{M} I[s_i^{t-1} > 0] < R_{keep} \times M$ **do**
5: $\{(A_i, I_i, y_i)\}_{i=1}^{N} \leftarrow$`SampleMiniBatch`$(\mathcal{D}, N)$
6: $\boldsymbol{x}_i, \boldsymbol{v}_i \leftarrow F_{\theta_f}(I_i), G_{\theta_g}(A_i)$.
7: `Calculate` $\mathcal{L}_{implicit}$ `with` Eq. (1) `and` (2).
8: `Update` H^t `with` Eq. (6).
9: **if** $t \% T_{update} = 0$ **then**
10: `Update` s^t `with` Eq. (7).
11: **else**
12: $s^t \leftarrow s^{t-1}$.
13: **end if**
14: `Update` θ_f, θ_g `with` Eq. (11).
15: **end while**
16: ▷ `Third Stage`
17: `Reinitialize` θ_f, θ_g.
18: **for** $t = 1$ to T_{max} **do**
19: `Update` θ_f, θ_g `with` Eq. (11).
20: **end for**

The details for the first and third stage are obvious. We only elaborate on the second stage next.

During this training phase, we update the weights in an iterative manner. In the t-th iteration, we first sample a mini-batch (`Line 5` in Alg. 1) and perform a standard inference to obtain the embeddings $\boldsymbol{x}_i, \boldsymbol{v}_i$ for all the instances in the mini-batch (`Line 6` in Alg. 1). Next, we evaluate the hardness of each involved identity in the current mini-batch. To ensure the robustness of the learned weights, we only employ the more stable implicit loss (`Line 7` in Alg. 1) $\mathcal{L}_{implicit}$ to measure the hardness. To integrate the hardness information from previous iterations and the current iteration, we perform a moving average strategy to calculate the hardness (`Line 8` in Alg. 1). Above all, the hardness for the i-th identity, *i.e.*, H_i^t, could be written as:

$$H_i^t = \beta \cdot H_i^{t-1} + (1 - \beta) \cdot \mathcal{L}_{implicit} \tag{6}$$

where $\beta \in (0, 1)$ is a hyper parameter to control the importance of the current iteration.

Now, we assign a weight for each identity to gradually add the hard identities to the training set (`Line 11` in Alg. 1). Specifically, in the t iteration of the training process, the weight of the samples belonging to the i-th identity in the loss is denoted as s_i^t. As for the initialization, s_i^0 is set to 1 for the bottom 30% H_i^0, and is set to 0 otherwise. To maintain stability, the weight update is triggered after every T_{update} iterations (`Line 9` in Alg. 1). If the update is triggered, s_i^t will be updated as follows:

$$s_i^t = \begin{cases} 1, \textit{for all bottom } k\ H_i^t \textit{ with } s_i^{t-1} = 0. \\ \alpha \cdot s_i^{t-1}, \textit{otherwise.} \end{cases} \tag{7}$$

where $\alpha \in (0, 1)$ and k (a positive integer) are hyper parameters. Next, we explain how this equation works. First, we will pick out all the identities with the smallest k H_i^t and set their s_i^t to 1. For the rest of the identities, we simply employ a weight decay strategy by setting $s_i^t = \alpha \cdot s_i^{t-1}$. Meanwhile, the identities with the largest H_i^t are considered as personalized identities, which should be dropped out via a zero weight. To this end, we update the identity weights sequentially. We will terminate the second stage of training immediately after R_{keep} percent of identities are assigned with non-zero weights (`Line 4` in Alg. 1).

At the end of each iteration (if not terminated), the network parameters θ_f, θ_g are updated with weighted loss functions (`Line 14` in Alg. 1):

$$\mathcal{L}_{face} = -\sum_{i=1}^{N} \hat{s}_{y_i}^t \log \frac{\exp(\boldsymbol{\omega}_{y_i}\boldsymbol{x}_i)}{\sum_{j=1}^{M}\exp(\boldsymbol{\omega}_j\boldsymbol{x}_i)} \tag{8}$$

$$\mathcal{L}_{voice} = -\sum_{i=1}^{N} \hat{s}_{y_i}^t \log \frac{\exp(\boldsymbol{\omega}_{y_i}\boldsymbol{v}_i)}{\sum_{j=1}^{M}\exp(\boldsymbol{\omega}_j\boldsymbol{v}_i)} \tag{9}$$

$$\mathcal{L}_{explicit} = \sum_{i=1}^{N} \hat{s}^t_{y_i} \log(m + \frac{\sum\limits_{y_j \neq y_i} \exp(\boldsymbol{v}_i \hat{\boldsymbol{x}}_j)}{\exp(\boldsymbol{v}_i \hat{\boldsymbol{x}}_i)}) + \sum_{i=1}^{N} \hat{s}^t_{y_i} \log(m + \frac{\sum\limits_{y_j \neq y_i} \exp(\boldsymbol{x}_i \hat{\boldsymbol{v}}_j)}{\exp(\boldsymbol{x}_i \hat{\boldsymbol{v}}_i)}) \quad (10)$$

$$\mathcal{L}_\theta = \mathcal{L}_{face} + \mathcal{L}_{voice} + \mathcal{L}_{explicit} \quad (11)$$

where $\hat{s}^t_i = s^t_i / \sum\limits_{j=1}^{N} s^t_j$.

4. Experiments

4.1. Datasets

Following previous work [30, 27], we evaluate the proposed method on a constructed dataset based on VoxCeleb [18] and VGGFace [20] datasets. VoxCeleb is an audio-visual dataset of short human speech videos, which provides audios for our experiments. Meanwhile, VGGFace is a human face dataset of 2,622 identities. Since in our settings, it is not necessary to capture voices and faces from the same video, we use still faces from VGGFace instead of extracted faces from VoxCeleb. The intersection of VoxCeleb and VGGFace contains 1,225 identities after filtering low quality data. Then we split the data into train/validation/test sets without identity overlapping, and generate the queries for validation and test according to our evaluation protocols (See Sec. 4.2). The statistics of the resulted dataset are reported in Tab. 1.

Table 1. Details on the datasets.

	train	validation	test	total
audio clips	113,322	14,182	21,850	149,354
face images	104,724	12,260	20,076	137,060
identities	924	112	189	1,225
queries (V-F)	–	42,546	65,550	108,096
queries (F-V)	–	36,780	60,228	97,008

4.2. Implementation details

Network architecture. The face feature extractor is implemented with SE-ResNet-50 [9], which is a powerful backbone on multiple tasks of computer vision. The input is a face image of size $112 \times 112 \times 3$, which is normalized to $[-1, 1]$ by subtracting 127.5 and dividing 127.5, and the output is a 128-dimensional face embedding. The voice feature extractor is implemented with Thin-ResNet-34 [1, 35], which inputs a spectrogram of voice and outputs a 128-dimensional voice embedding. Spectrograms have 257 channels (256 frequency components and 1 DC component), which are generated with a hamming sliding window of width 25ms and of hop 10ms. The two networks are pre-trained with the face recognition task on MS-1M [6], and the audio speaker recognition on VoxCeleb2 [4], respectively.

Training strategy. The training process is split into three stages: warm-up training, identity weight learning and retraining with fixed identity weights, as mentioned in Sec. 3.3. In order to balance the number of samples with different identities, we sample several identities in an iteration, and then randomly sample one face image and one voice audio clip for each identity. For data pre-processing, we apply data augmentation including random rotation (from $-15°$ to $15°$), random cropping (112×112) for images, and random cropping in time axis (from 2.5 seconds to 5.0 seconds) for audios. In our model, the hyperparameters are set as follows: $m = 3.4$ (Eq. (9)), $\beta = 0.9$ (Eq. (6)), $\alpha = 0.99, k = 22$ (Eq. (7)), $T_{warm} = 500, T_{update} = 100, R_{keep} = 0.9$ (Alg. 1). We adopt the stochastic gradient descent (SGD) optimizer, where batch size and momentum are set to 64 and 0.9, respectively. The learning rate is initialized as 10^{-2}, and decays by 0.1 in the 2k and 3k iterations. The max iteration T_{max} is 10k, and the best model on the validation set is preserved for evaluation.

Evaluation protocol. To show the overall capacity of the proposed model in voice-face association learning task, we conduct the evaluation under the following four settings:

(a) **1:2 matching**. Given an instance from one modality as the probe, and two candidates from the other modality (including one from the same identity as the probe) as the gallery, the task is to find out which candidate matches the probe. The performance is measured with accuracy (ACC).

(b) **1:N matching**. This task is basically the same as 1:2 matching, except that the length of gallery N ranges from 2 to 10 in our experiments. We report ACC on each N.

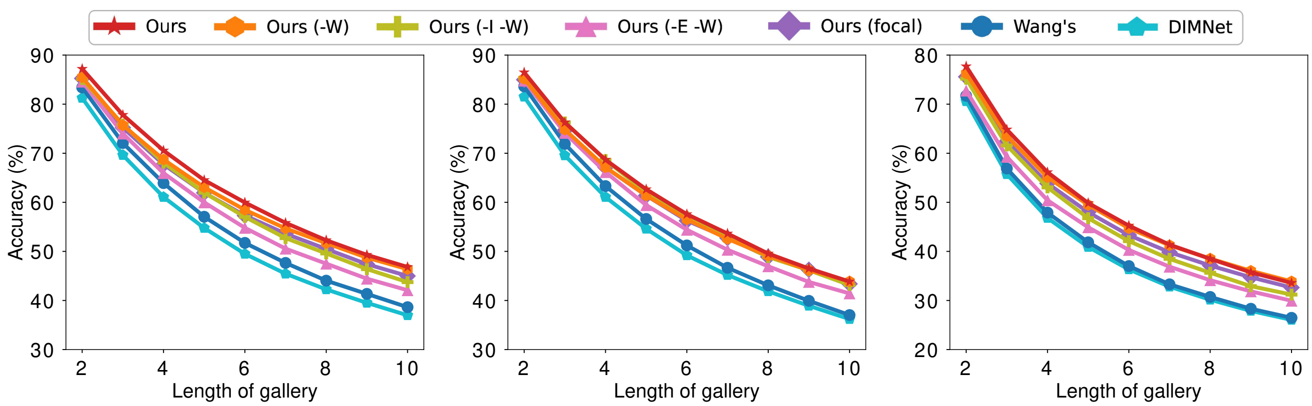
(c) **Verification**. Given two instances from different modalities, the task is to judge whether they belong to the same person. The performance is measured with Area Under the ROC curve (AUC).

(d) **Retrieval**. This task is extended from the 1:N matching task, where the gallery contains one or more candidates that match the probe. The model is asked to rank the gallery samples so that the candidates matching the probe are ranked at the top. We report the performance with mean average precision (mAP).

For each task, we report the metrics for two types of queries, from voice to face (**V-F**) and from face to voice (**F-V**). On matching and verification tasks, the queries are further divided into two subtypes: gallery samples have the same gender as the probe sample (**G**), or have unrestricted genders (**U**).

Table 2. Results (%) on 1:2 matching, verification and retrieval. V-F: from voice to face; F-V: from face to voice; U: unrestricted; G: query restricted by gender. The best results of our models and competitors are highlighted in soft red and soft blue, respectively.

Methods	1:2 Matching (ACC)				Verification (AUC)				Retrieval (mAP)	
	V-F (U)	F-V (U)	V-F (G)	F-V (G)	V-F (U)	F-V (U)	V-F (G)	F-V (G)	V-F	F-V
SVHF [17]	81.0	79.5	63.9	63.4	–	–	–	–	–	–
DIMNet [30]	81.3	81.9	70.6	69.9	81.0	81.2	**70.4**	69.3	4.3	**3.8**
Wang's [27]	**83.4**	**84.2**	**71.7**	**71.1**	**82.6**	**82.9**	70.3	**70.1**	**4.4**	3.4
Ours (focal [14])	85.3	85.0	75.6	74.5	85.6	85.4	76.0	75.1	6.5	6.2
Ours (-E -W)	84.6	84.7	72.8	71.3	84.8	85.0	72.4	71.4	4.6	5.1
Ours (-E)	84.3	84.3	75.0	74.4	84.6	84.6	75.3	74.8	5.1	4.9
Ours (-I -W)	85.1	85.4	75.2	74.3	85.7	85.7	75.8	75.2	4.8	5.1
Ours (-W)	85.5	85.2	76.3	74.7	85.8	85.3	76.7	75.1	**6.5**	**6.2**
Ours	**87.2**	**86.5**	**77.7**	**75.3**	**87.2**	**87.0**	**77.5**	**76.1**	5.5	5.8



(a) V-F (U) (b) F-V (U) (c) V-F (G)

Figure 3. Quantitative results on 1:N matching task. Best viewed in color.

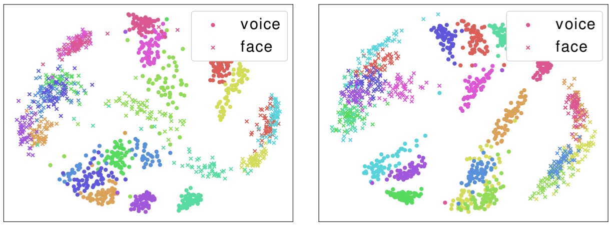
Competitors. We compare our proposed method with three models: SVHF [17], DIMNet [30] and Wang's model [27]. Note that SVHF takes dynamic faces as input, so its dataset is not exactly the same as ours.

Ablated variants. In order to quantify two main contributions of this paper: two-level modality alignment and identity re-weighting, we also implement four ablated variants: (1) the model without explicit alignment and re-weighting (denoted by **-E -W**), (2)the model without implicit alignment and re-weighting (denoted by **-I -W**) (3) the model without re-weighting (denoted by **-W**), and (4) a model trained with focal loss [14] instead of our re-weighting strategy (denoted by **focal**).

4.3. Results and comparison

Quantitative results. The results on the 1:2 matching, verification and retrieval are recorded in Tab. 2. We see that our model significantly outperforms the competitors over all three tasks, with an average improvement of about 2%-7%. Furthermore, comparing the performance on gender restricted and unrestricted queries, our method has a larger improvement in the case of gender restriction. These validate that our model could discover deeper associations between face and voice. On the other hand, the results of 1:N matching are shown in Fig. 3, which further verifies the advantage of our method. In this task, the accuracy decreases with the increase of N, but our method consistently has a higher performance and has less decrease than the competitors, which shows that the proposed framework is relatively more robust.

Moreover, we could make the following observations: (1) The explicit modality alignment can significantly improve performance, especially in the gender-restricted groups, which are more challenging. In the 1:2 matching task and the verification task, identity re-weighting brings an increase of about 1%, and achieves the state-of-the-art performance. However, training with focal loss doesn't bring significant improvement. (2) For all the tasks, results on gender-unrestricted queries are obviously better than gender-restricted ones. This shows that facial and voice features are strongly related to gender, which is consistent with people's experience. Besides, even if the gender is re-



(a) Embeddings on training set. (b) Embeddings on test set.

Figure 4. Visualization of learned embeddings. The same color points belong to the same person, and different shapes represent different modalities. Dimension reduction is implemented by multi-dimensional scaling (MDS) [32]. Best viewed in color.

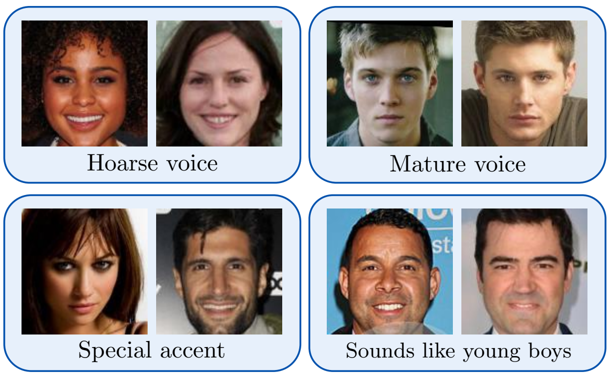


Figure 5. Some examples of personalized samples. We show four cases of voice and face inconsistency as examples.

stricted, machines can still learn the association between face and voice. (3) Despite that our identity re-weighting strategy brings improvements on the matching and verification tasks, the full model does not perform as well as its ablated counterpart on the retrieval task. One possible reason is that the re-weighting strategy improves generalization at the expense of precision on the tail of gallery. How to improve the performance of this situation more effectively is worthy of further study.

Qualitative results. A visualization of the face embeddings and the voice embeddings extracted with our model is provided in Fig. 4. It could be observed that embeddings from the same identities are close and discriminative in most cases, which shows the effectiveness of learned features intuitively.

4.4. Ablation study for training strategy

Effect of identity weights. In our framework, the hard identities are filtered out by manually setting the ratio of data retained R_{keep}. As is shown in Tab. 3, we evaluate the 1:2 matching accuracy of models learned with different R_{keep}, ranging from 0.6 to 1.0. Three conclusions can be drawn: (1) Weighting the difficult samples can bring about 1.2% to 1.6% performance improvement, even if ex-

Table 3. The effect of weighting and personalized identities filtering on 1:2 matching accuracy. The best results are shown in **bold**.

R_{keep}	re-weighting	V-F (U)	F-V (U)
1.0	×	85.5	85.2
1.0	✓	86.7	86.8
0.9	✓	**87.2**	**87.2**
0.8	✓	87.0	87.1
0.7	✓	86.3	86.3
0.6	✓	85.9	85.3

tremely hard identities are not excluded. (2) After dropping extremely hard identities, the performance is further boosted, verifying the effectiveness of eliminating personalized identities. (3) After a certain point, the performance of the model decreases with the increase of the retaining ratio. This is due to the abandonment of some data might lower the data diversity. Some excluded samples are shown in Fig. 5. Intuitively speaking, these samples are usually special in voice. See *Supplementary Materials* for more analysis.

Effect of pre-training. We also explored the effect of pre-training on the model in *Supplementary Materials*.

5. Conclusion

In this paper, we propose a novel embedding framework for voice-face assoication learning, with **(a)** a comprehensive modality alignment loss embracing global and local information; **(b)** a dynamic re-weighting strategy to deal with the diversity of learning difficulty across identities. For **(a)**, we propose a two-level loss including implicit modality alignment loss and explicit modality alignment loss. In Prop. 1, we prove that the implicit alignment loss could globally reduce the inconsistency of the embeddings across modality. For **(b)**, we propose a re-weighting framework which can focus on hard identities while filter out personalized identities via the evolution of the identity weights. Experiments on cross-modal matching, verification and retrieval show that compared with previous methods, our method can better learn the association between face and voice.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102003, in part by National Natural Science Foundation of China: 61620106009, 61931008, 61836002, and 61976202, in part by Youth Innovation Promotion Association CAS, and in part by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB28000000.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 6

[2] Yeqi Bai, Tao Ma, Lipo Wang, and Zhenjie Zhang. Speech fusion to face: Bridging the gap between human's vocal characteristics and facial imaging. *arXiv preprint arXiv:2006.05888*, 2020. 1

[3] Hyeong-Seok Choi, Changdae Park, and Kyogu Lee. From inference to generation: End-to-end fully self-supervised generation of human face from speech. In *International Conference on Learning Representations*, 2019. 1

[4] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *Proc. Interspeech*, pages 1086–1090, 2018. 6

[5] Yang Fan, Fei Tian, Tao Qin, Jiang Bian, and Tie-Yan Liu. Learning what data to learn. *arXiv preprint arXiv:1702.08635*, 2017. 3

[6] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102, 2016. 6

[7] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742, 2006. 2

[8] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. Face-voice matching using cross-modal embeddings. In *ACM International Conference on Multimedia*, pages 1011–1019, 2018. 1, 3

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6

[10] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. Putting the face to the voice': Matching identity across modality. *Current Biology*, 13(19):1709–1714, 2003. 1

[11] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *Asian Conference on Computer Vision*, pages 276–292, 2018. 1, 2

[12] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017. 3

[13] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 8577–8584, 2019. 3

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 3, 7

[15] Lauren W Mavica and Elan Barenholtz. Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2):307, 2013. 1

[16] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *European Conference on Computer Vision*, pages 71–88, 2018. 1, 2

[17] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8427–8436, 2018. 1, 2, 7

[18] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *Telephony*, 3:33–039, 2017. 6

[19] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7539–7548, 2019. 1

[20] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, pages 41.1–41.12, September 2015. 6

[21] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343, 2018. 3

[22] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016. 3

[23] Harriet MJ Smith, Andrew K Dunn, Thom Baguley, and Paula C Stacey. Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, 78(3):868–879, 2016. 1

[24] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016. 3, 4

[25] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 4

[26] Maria Stephanie Tsantani. *Brain and perceptual representations of faces, voices, and person identity*. PhD thesis, Brunel University London, 2019. 1

[27] Rui Wang, Xin Liu, Yiu-ming Cheung, Kai Cheng, Nannan Wang, and Wentao Fan. Learning discriminative joint embeddings for efficient face and voice association. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1881–1884, 2020. 1, 3, 6, 7

[28] Tianyang Wang, Jun Huan, and Bo Li. Data dropout: Optimizing training data for convolutional neural networks. In *IEEE International Conference on Tools with Artificial Intelligence*, pages 39–46, 2018. 3

[29] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1473–1480, 2006. 2

[30] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh. Disjoint mapping network for cross-modal matching of voices and faces. In *International Conference on Learning Representations*, 2018. 1, 3, 6, 7

[31] Yandong Wen, Bhiksha Raj, and Rita Singh. Face reconstruction from voice using generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 5265–5274, 2019. 1

[32] Florian Wickelmaier. An introduction to mds. *Sound Quality Research Unit, Aalborg University, Denmark*, 46(5):1–26, 2003. 8

[33] Chuyuan Xiong, Deyuan Zhang, Tao Liu, and Xiaoyong Du. Voice-face cross-modal matching and retrieval: A benchmark. *arXiv preprint arXiv:1911.09338*, 2019. 1, 2

[34] Y Zhang, S Yang, J Xiao, S Shan, and X Chen. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 851–858, 2020. 1

[35] Yujie Zhong, Relja Arandjelović, and Andrew Zisserman. Ghostvlad for set-based face recognition. In *Asian Conference on Computer Vision*, pages 35–50, 2018. 6