Hearing like Seeing: Improving Voice-Face Interactions and Associations via Adversarial Deep Semantic Matching Network

Kai Cheng[1,2], Xin Liu[1,2,3,*], Yiu-ming Cheung[3,*], Rui Wang[1], Xing Xu[4], Bineng Zhong[1,5]
[1] Department of Computer Science, Huaqiao University, Xiamen 361021, China
[2] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[3] Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong
[4] School of Computer Science and Engineering, University of Electronic Science and Technology of China
[5] Xiamen Key Lab. of Computer Vision and Pattern Recognition, Fujian Key Lab. of Big Data Intelligence and Security
{kcheng, rw}@stu.hqu.edu.cn, {xliu, bnzhong}@hqu.edu.cn, ymc@comp.hkbu.edu.hk, xing.xu@uestc.edu.cn
*Corresponding authors.

ABSTRACT

Many cognitive researches have shown that human may 'see voices' or 'hear faces', and such ability can be potentially associated by machine vision and intelligence. However, this research is still under early stage. In this paper, we present a novel adversarial deep semantic matching network for efficient voice-face interactions and associations, which can well learn the correspondence between voices and faces for various cross-modal matching and retrieval tasks. Within the proposed framework, we exploit a simple and efficient adversarial learning architecture to learn the cross-modal embeddings between faces and voices, which consists of two subnetworks, respectively, for generator and discriminator. The former subnetwork is designed to adaptively discriminate the high-level semantical features between voices and faces, in which the triplet loss and multi-modal center loss are in tandem utilized to explicitly regularize the correspondences among them. The latter subnetwork is further leveraged to maximally bridge the semantic gap between the representations of voice and face data, featuring on maintaining the semantic consistency. Through the joint exploitation of the above, the proposed framework can well push representations of voice-face data from the same person closer while pulling those representations of different person away. Extensive experiments empirically show that the proposed approach involves fewer parameters and calculations, adapts various cross-modal matching tasks for voice-face data and brings substantial improvements over the state-of-the-art methods.

CCS CONCEPTS

• **Information systems → Novelty in information retrieval**; *Information retrieval*;

KEYWORDS

Voice-face association, adversarial deep semantic matching, multi-modal center loss, cross-modal embeddings

ACM Reference Format:
Kai Cheng[1,2], Xin Liu[1,2,3,*], Yiu-ming Cheung[3,*], Rui Wang[1], Xing Xu[4], Bineng Zhong[1,5]. 2020. Hearing like Seeing: Improving Voice-Face Interactions and Associations via Adversarial Deep Semantic Matching Network. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394171.3413710

1 INTRODUCTION

Many researches in cognitive science and neuroscience have shown that humans often integrate audio-visual information for various perception tasks [5, 12]. In particular, face and voice cues, incorporating the advantages of non-intrusiveness and easily accessible, have been extensively utilized for various identity recognition tasks [24]. As humans, we can recognize the identity of a familiar person from either face or voice alone. In recent years, some cognitive researches have also confirmed that humans are able to hear voices of known individuals to form mental pictures of their facial appearances, *i.e.*, hearing faces, while memorizing and recalling voices when seeing to their facial pictures, *i.e.*, seeing voices [8, 9]. This is practical feasible, *e.g.*, we are able to draw a general picture of a speaking face on the other end the line in practice. Even, these observations also explicitly provide strong support that humans are capable of correctly matching unfamiliar face images to the corresponding voice recordings, and vice versa.

From another perspective, the above studies lend credence to the hypothesis that it may be possible to emulate the human ability to find the associations between voices and faces intelligently. Automatic voice-face association is particularly useful for creating natural human machine interaction systems and benefiting many valuable applications such as speaker annotation and diarization. For instance, a media recommender system with voice-face association embedding can well present an appropriate voice actor for a cartoon character, and this phenomenon can be defined as voice-face matching task. With the motivation of whether machine learning models can reveal the correlations across different media data, there have emerged some studies to mine the correlations between audio and visual examples [10, 13]. Inspired by recent advance of multi-modal deep learning works, some approaches [11, 19] generally correlate the faces and voices by having the same identity, and leverage deep neural networks to exploit the common embeddings from paired voice-face examples. It is noted that these common embedding models are likely to succeed on the relevant training

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413710

data, but which often induce poor performances on the unrelated data. To the best of our knowledge, researches on reliable voice-face association problem are still under the early stage.

In this paper, we focus on strengthening the semantic association between the face examples and relevant voice recordings, while developing an intelligent algorithm to predict a voice from faces and a face from voice clips. To this end, we develop an efficient adversarial deep semantic matching (ADSM) network, consisting of generator and discriminator networks, to tackle voice-face association problem. As shown in Figure 1, on the one hand, we develop an efficient generative subnetwork to learn the high-level semantical features between voice and face, and sequentially utilize the triplet loss and multi-modal center loss to explicitly regularize the correspondences among voice-face data. On the other hand, we propose to utilize adversarial subnetwork to maximally reduce the semantic gap between the representations of voice and face data, which can well preserve modality similarity while minimizing inter-modal ambiguity. Through the joint exploitation of the above, the proposed ADSM framework can well push representations of voice-face data from the same person closer while pulling those representations of different person away. Extensive experiments show the promising results on various cross-modal matching tasks, including selection of a face or voice from a pair of choices, verification a given pair of instances belong to the same person, and retrieval the given input of one modality to one-out-of-N matching. The major contributions of this paper are highlighted bellow:

- A novel ADSM framework is efficiently proposed for reliable voice-face interactions and associations, which can well facilitate various cross-modal matching tasks.
- An efficient multi-modal center loss is presented for cross-model common embedding learning, which can explicitly regularize the correspondences among the voice-face data.
- A discriminative adversarial learning mechanism is addressed to seamlessly guide the high-level semantical feature learning and bridge the semantic gap between the heterogeneous voice and face data.
- Extensive experiments demonstrate the advantages of the proposed framework under various cross-modal matching tasks, and show its outstanding performance in comparison with the-state-of-art methods.

The rest of this paper is organized as follows: Section 2 briefly surveys the voice-face association works, and Section 3 elaborates the proposed model and its implementation details. The extensive experiments and comparisons are introduced in Section 4. Finally, we draw a conclusion in Section 5.

2 RELATED WORKS

Human may hear faces by forming mental pictures of what a person looks like after only hearing his or her voice, while recalling voices when taking a glance at their face pictures [24]. This phenomenon has been investigated in a number of studies on human perception and neurology [8, 12, 30]. These cognitive studies lend credence to prove that it may be possible to find associations between voices and faces, and machine learning community has also shown increasing interest in studying such associations. Alone this line, some works [4, 15] develop systems to identify active speakers from a video by jointly observing the audio and visual signals. Although these voice-speaker matching task seems similar, they mainly focus on distinguishing active speakers from non-speakers at a given time. By assuming that the voice and face are implicitly captured from one speaker, these works learn the common embeddings through joint presentation of voices and faces, to maximize their similarities if they belong to the same speaker. Within these works, the voice data and face image are acquired simultaneously.

With the recent advance of deep learning, multi-modal deep learning leverages neural networks to mine common information from large-scale multi-modal media data [21]. Inspired by these works, SVHF [19] utilizes CNN architectures to learn the joint presentation of voices and faces, and formulates their association problem as a binary selection task. Later, they further form the positive voice-face pairs acquired from the same talking face in a video, and consider the negative voice-faces pairs from different videos. Accordingly, they utilize the contrastive loss to minimize the distance between the embeddings of positive pairs and penalizes the negative pair distances [18]. Experimentally, these models are able to match human performance on some challenging examples, *e.g.*, faces with the same gender, age and nationality. Nevertheless, they are task-specific methods, which cannot be utilized for other matching applications. Similarly, FV-CME [6] first exploits a face-voice matching model to learn cross-modal embeddings, and further utilizes the N-pair loss to regularize the correspondence in voice-face feature learning process. This approach is able to achieve face-voice matching task, but which needs the fine-tuning process to optimize the model, whose parameters is very huge. In addition, LAFV [11] exploits the overlapping information between faces and voices, and utilizes the standard network architectures to learn their common latent spaces. Without human supervision, this approach trains the networks from naturally paired face-voice data, and yields similar results have been reported by other researches [6, 19]. Differently, DIMNet [25] does not explicitly learn the joint relationship between different modalities, but learns a shared representation by mapping them individually to their common covariates, *i.e.*, identity, nationality and gender. This work is designed to be data less-intensive, but which does not fully consider the high-level semantic correlations between different modalities and also involve large network parameters for learning. Therefore, its performances need further improvements.

3 PROPOSED APPROACH

3.1 Problem Definition

The main objective of this paper is to study the cross-modal associations between face and voice. Without loss of generality, we first tackle this problem as a binary matching task. Let $\mathcal{F}$ represent the face samples, $\mathcal{V}$ denote the voice data, $\mathcal{A}$ characterize the attributes of face and voice, *i.e.*, $\mathcal{A}_{ID}(\cdot)$: **Identity**; $\mathcal{A}_{G}(\cdot)$: **Gender**; $\mathcal{A}_{N}(\cdot)$: **Nationality**. Accordingly, the face-voice modality dataset can represented as $\mathcal{M} = \{\mathcal{F}, \mathcal{V}\}$. For voice to face matching task (V⟶F) that consisting of a voice clip and two face images, the goal is to examine which face is more likely to associate the voice clip. Suppose there exist a triplet data set $\mathbf{x}$=$\{\mathbf{v}, \mathbf{f}_1, \mathbf{f}_2\}$ consisting of an anchor voice segment clip $\mathbf{v}\in\mathcal{V}$ and two face images $\mathbf{f}_1\in\mathcal{F}$ and $\mathbf{f}_2\in\mathcal{F}$. Note that, face examples in $\mathbf{x}$ contain a positive sample and a

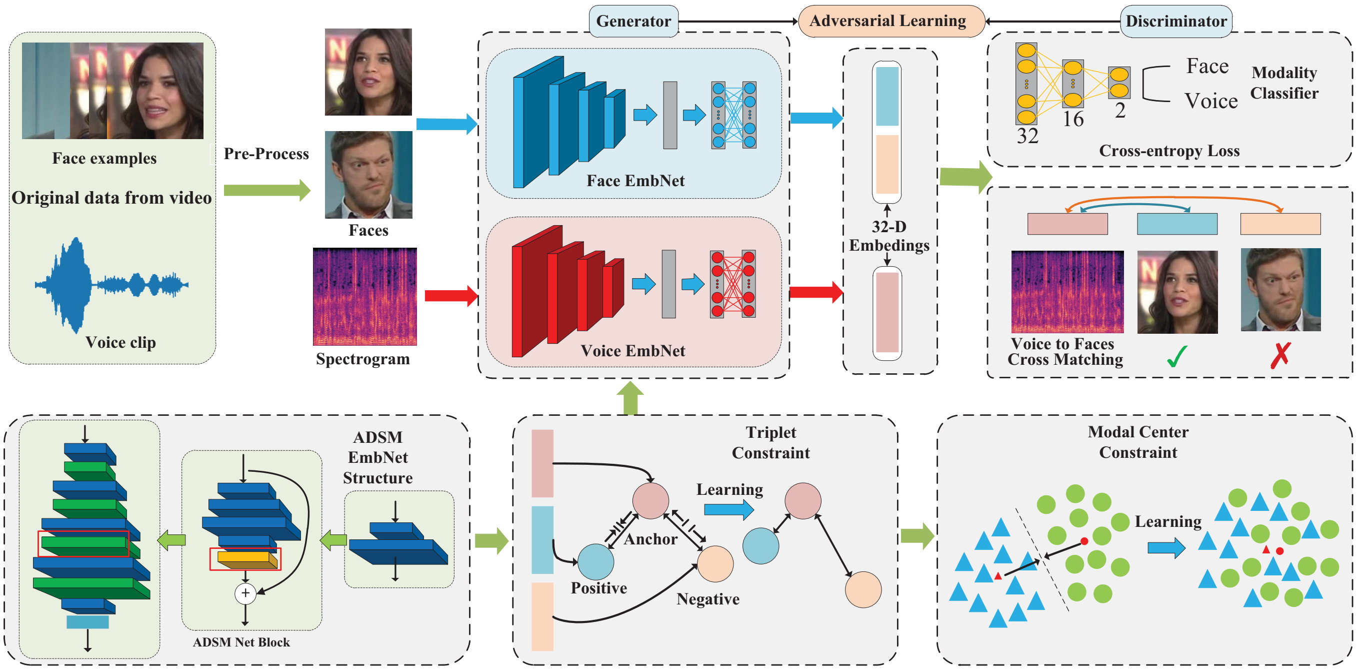

Figure 1: The schematic pipeline of the proposed ADSM network with adversarial training architectures.

negative sample. More specifically, $\mathbf{f}_i$ can be considered as a positive example if it possesses the same attributes as the anchor voice, *i.e.*, $\mathcal{A}_{ID}(\mathbf{f}_i)$=$\mathcal{A}_{ID}(\mathbf{v})$, and vice versa. Therefore, such voice-face association task can be considered as a binary classification task, which can be achieved by predicting the position $y\in\{0,1\}$ of the positive face. Meanwhile, let v=$N_\mathbf{v}(\theta_\mathbf{v};\mathbf{v}) \in \mathbb{R}^d$ and f=$N_\mathbf{f}(\theta_\mathbf{f};\mathbf{f}) \in \mathbb{R}^d$ respectively represent the voice mapping function and face mapping function, which can individually map voice feature vector and face feature vector into d-dimensional common subspace, and $\theta_\mathbf{v}$ and $\theta_\mathbf{f}$ are their function parameters. Then let the $m = N_\mathbf{d}(\theta_\mathbf{d}; x)$ represent the discriminator in ADSM framework, where m is the output of modality classifier in discriminator and x is a d-dimensional vector that obtained from the output of $N_\mathbf{f}$ or $N_\mathbf{v}$, and $\theta_\mathbf{d}$ is the parameter for modality classifier in discriminator.

3.2 Proposed Framework

As shown in Figure 1, we exploit an efficient adversarial deep semantic matching (ADSM) network for voice-face association, which of two subnetworks, respectively, for generator and discriminator. The former subnetwork is designed to adaptively discriminate the high-level semantical features between voices and faces, in which the triplet loss and multi-modal center loss are in tandem utilized to explicitly regularize the correspondences among them. The latter subnetwork is further leveraged to maximally reduce the semantic gap between the representations of voice and face data, and modality classifier is utilized as discriminator **G** to maintain the semantic consistency among them. By acting as an adversary, the joint adversarial loss is further utilized for narrowing the gap between voice and face data, by learning as a minimax game. As a whole, these two subnetworks are trained together in an end-to-end manner.

3.3 Network Model

Within the proposed learning model, the main purpose of the generative subnetwork is to extract the discriminative high-level features that can well characterize the voice-face data, while maintaining the semantic correspondences between them. Meanwhile, it is reasonable to expect that the learning network should be more powerful with fewer parameters and can be calculated quickly [3]. In recent years, ResNeXt [26] is demonstrated to be a simple and highly modularized network architecture for image classification. More specifically, this model repeats a building block that aggregates a set of transformations with the same topology, which involves less hyper-parameters while gaining accuracy. Meanwhile, Squeeze-and-Excitation (SE) [7] block adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels, which brings significant performance improvements at slight additional computational cost. Inspired by these findings, we propose to stack ResNeXt and SE blocks together to emphasize informative visual features and suppress less useful ones, in which the network configuration is shown in Table 1.

The proposed subnetwork is shown in the lower part of Figure 1, where the green layer is convolutional layer, the blue layer is ADSM block and the yellow layer is SE module. The output dimension of every convolutional layer situated in front of block is twice than the input dimension, and top three layers in SE-ResNeXt block double the dimension. Meanwhile, the last convolutional layer will halve the dimensions before the feature embedded into the SE module.

Table 1: The network configurations of face embnet.

Name	Channel Num	Output Size	Stride
Input	3	128×128	-
Conv1	64	64×64	2
ADSM block1	64	64×64	-
Conv2	128	32×32	2
ADSM block2	128	32×32	-
Conv3	256	16×16	2
ADSM block3	256	16×16	-
Conv4	512	8×8	2
ADSM block4	512	8×8	-
Conv5	32	4×4	2
Global pool	32	-	-

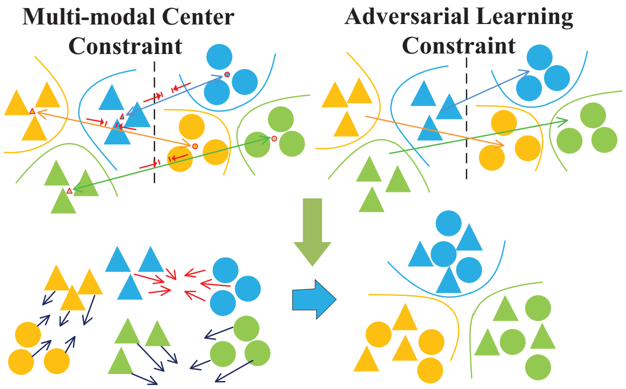
Multi-modal Center Constraint
Adversarial Learning Constraint

Figure 2: Illustration of MMC and adversarial constraints.

3.4 Loss Function

Face and voice data often exhibit diverse representations, and we first utilize the triplet loss to regularize the correspondences between the face and voice. Since there are many categories in the dataset, it is difficult to increase the distance between different classes. Therefore, we expect the distance within the same category to be as small as possible, and the triplet loss is employed:

$$\mathcal{L}_{Tri} = \sum_{i=1}^{n}[d(\mathbf{a},\mathbf{p}) - \alpha d(\mathbf{a},\mathbf{e}) + \text{margin}]_{+} \tag{1}$$

where **a** is anchor, **p** is positive sample, **e** is negative sample, margin is a constant that utilized to maintain the positive distance value, and $d(\cdot,\cdot)$ is Euclidean distance, hyper parameter α is utilized to control weights between the positive distance and negative distance. More specifically, in V→F task, **a**, **p** and **e** respectively correspond to voice, positive face and negative face, and vice versa.

The goal of the proposed framework is to minimize the semantic gap among the high-level representations of semantically similar voice-face samples. The preservation of intra-modal similarity ensures that the representations of data points with the same semantical category should be close to each other, while the preservation of inter-modal similarity is often utilized to maintain the semantic consistency between heterogeneous modalities [1, 2, 16, 29]. As the triplet loss converges, then we change the triplet loss to a novel multi-modal center (**MMC**) loss. As shown in Figure 2, we add a constraint between features and their centers, which contains two part: intra-modality center distance and inter-modality center distance. Intra-modality center distance makes the data distribution within the same modality to be much closer, while the inter-modality center distance ensures the centers of heterogeneous modalities within the same category to be as close as possible. Before introducing the MMC loss, the centers for each class and modality are calculated:

$$c_i^j = \frac{1}{n}\sum_{k=1}^{n} x_k,\ c_i = \frac{1}{m}\sum_{k=1}^{m} c_i^k \tag{2}$$

where c_i^j means the j-th modality center in i-th class, c_i means the center of i-th class with all modalities, n, m are respectively the dimensionality of features and the number of categories. Accordingly, the MMC loss can be formulated as follows:

$$\mathcal{L}_{MMC-intra} = \frac{1}{2}\sum_{i=1}^{n}\left\|N(x_i^j) - c_i^j\right\|_2^2 \tag{3}$$

$$\mathcal{L}_{MMC-inter} = \frac{1}{2}\sum_{i=1}^{n}\|N(x_i) - c_i\|_2^2 \tag{4}$$

Then, the integrated MMC loss can be formulated as:

$$\mathcal{L}_{MMC} = \mathcal{L}_{MMC-intra} + \gamma\mathcal{L}_{MMC-inter} \tag{5}$$

where γ is the balance parameter. Accordingly, the generative loss for generator subnetwork is given by:

$$\mathcal{L}_{\mathbf{g}} = \begin{cases} \mathcal{L}_{Tri}, & epoch < k \\ \mathcal{L}_{MMC}, & epoch \geq k \end{cases} \tag{6}$$

For adversarial learning, the modality classifier $N_{\mathbf{d}}$ is utilized as a discriminator, which is defined to discriminate the data from the face to voice and it also acts as an adversary. As shown in Figure 2, it is expected the distances further regularized by the adversarial learning could ensure that the representations of data points with the same semantical category should be close to each other [14]. Within the proposed ADSM, modality classifier is acted as a three layer fully connected neural network and the cross-entropy loss is utilized to bridge the semantic gap between different modalities:

$$\mathcal{L}_{\mathbf{d}} = \frac{1}{m}\sum_{j=1}^{m}(j \cdot logN_{\mathbf{d}}(x)) \tag{7}$$

where x is input vector from the output of $N_{\mathbf{f}}$ or $N_{\mathbf{v}}$ and m is the modality number. During the training process, we need to learn both the generator and discriminator in an end-to-end way, and the optimization process contains two part:

$$\theta_{\mathbf{f}}, \theta_{\mathbf{v}} = \arg\min_{\theta_{\mathbf{f}},\theta_{\mathbf{v}}}(\mathcal{L}_{\mathbf{g}} - \eta\mathcal{L}_{\mathbf{d}}) \tag{8}$$

$$\theta_{\mathbf{d}} = \arg\max_{\theta_{\mathbf{d}}}(\mathcal{L}_{\mathbf{g}} - \eta\mathcal{L}_{\mathbf{d}}) \tag{9}$$

where η is a hyperparameter to control the training weight between discriminator and generator. This minimax game can be efficiently implemented using a stochastic gradient descent optimization solver, and the optimizations can be iteratively solved until the convergence is reached. In summary, the proposed learning algorithm for ADSM is displayed in **Algorithm** 1.

Algorithm 1 Learning algorithm of ADSM framework

```
Input: Face-voice Dataset: ${\mathbf{a}, \mathbf{p}, \mathbf{n}}$ corresponds to the original
    input data ${\mathbf{v}, \mathbf{f}_1, \mathbf{f}_2}$ or ${\mathbf{f}, \mathbf{v}_1, \mathbf{v}_2}$,
    Initialize face generator $f$=$N_f(\theta_f; \mathbf{f})$, voice generator
    $v$=$N_v(\theta_v; \mathbf{v})$,
    Initialize model discriminator by: $m = N_\mathbf{d}(\theta_\mathbf{d}; x)$,
    Initialize hyperparameters: $\gamma$, $\eta$, $k$, $H$
Repeat until convergence:
    if train epoch num == $k$ then
        for $i$ = 0 to $class\_num$ do
            $c_i$ = (mean($\Sigma N(a_i)$) + mean($\Sigma N(p_i)$))/2
        end for
        use set $c$ to initialize MMC loss
    end if Calculate the ADSMnet loss
    $a = N(\mathbf{a}), p = N(\mathbf{p}), n = N(\mathbf{n})$
    if train epoch num < $k$ then
        $\mathcal{L}_g = \mathcal{L}_{Triplet}(N(a), N(p), N(n))$
    else
        $\mathcal{L}_g = \mathcal{L}_{MMC}(N(a), N(p))$
    end if
    Calculate the Modality Classifier loss
    $p_1 = N_\mathbf{d}(a), p_2 = N_\mathbf{d}(p)$
    $\mathcal{L}_d = \mathcal{L}_{CE}(p_1) + \mathcal{L}_{CE}(p_2)$
    Train the network
    if train epoch num %$H$! =0 then
        $\mathcal{L}_{all} = \mathcal{L}_g - \eta\mathcal{L}_d$
    else
        $\mathcal{L}_{all} = \eta\mathcal{L}_d - \mathcal{L}_g$
    end if
    Return: $\mathcal{L}_{all}$
```

3.5 Model Training

The proposed ADSM networks are trained in an end-to-end manner by stochastic gradient descent (SDG) with batch normalization, and SGD optimizer with momentum 0.9, weight decay 0.001 and batch size 96 is selected for training on GPU NVIDIA RTX 2080Ti. For generator, the learning rate is initialized with 1e-5 at the beginning of training process, changed to 1e-2 when at the 6 epoch, divided by 10 after every 3 epochs, and changed to MMC loss when the loss function converges to a stable value. Then, the learning rate is reset to 10e-3 and divided by 10 after every 3 epochs, until the loss function converges. For discriminator, the learning rate is also set at 1e-4 for warming up, set to 1e-2 at 6 epoch, divided by 10 after every 4 epochs. In addition, training round H is set at 5, η is fixed to be 0.3, α is set at 0.8 and margin value is fixed at 0.6.

4 EXPERIMENTS

Dataset: The public VoxCeleb [20] and VGGFace [22] datasets are selected for evaluation, which consist of 1251 celebrities recorded under different environments and background noise levels. For fair comparison, we refer to the challenging split scheme [19], where persons of names starting with ['A', 'B'] are assigned to the validation set, while those starting with ['C', 'D', 'E'] are assigned for testing, and the rest persons are selected as the train set, summarized in Table 2. The proposed ADSM method trained through full learning scheme is abbreviated as **ADSM-TCD**. Meanwhile, the proposed network learned only use triplet loss is abbreviated as **ADSM-T** and use both triplet loss and MMC loss is abbreviated as **ADSM-TC**. Since the work [18] select 901 identities in train set and 250 identities in test set, which is different from the split from work [19]. Further, we also train our network through this data split with less training data, and name it as **ADSM-LTCD**.

4.1 Data Pre-Processing

Face examples: All the facial regions are detected by MTCNN [27], and the sizes of cropped RGB face examples are scaled to 128×128. Similar to [23, 25**?**], the face images are augmented by horizontally flipping the images with 50% probability, and normalized to $[-1, 1]$.
Voice clip: The speaking voice clips are detected by voice activate detector [28]. If the recording length of detected voice clip is less than 10s, we randomly repeat it to 10s to ensure the adequate samples. Conversely, we cut off the voice clip to be 10s if its length is larger than 10s. Consequently, we calculate the MFCC features by using the window 25ms and interval 10ms.

Table 2: Number statistics for the splitting datasets.

Item	Train	Validation	Test	Total
Identities	942	116	193	1251
Face Images	930428	112492	174512	1217432
Voice clips	116501	14635	22380	153516
Videos	16820	2044	3425	22295

4.2 Cross-modal Matching Protocol

In order to verify the effectiveness of the proposed ADSM approach, we design the following protocols for evaluation.
1:2 matching task: Each input data pair contains an Anchor from one modality (voice or face), and a gallery of two inputs from the other modality (face or voice), including a positive example that belongs to the same person and a negative example that does not match the person. The task of the network is to determine which sample is a positive example, and we have 22118400 groups of testing data in total.
1:N matching task: This task is the extension of 1:2 matching, and the model need to predict the only positive sample from N samples. This task is more challenging with the increase of number N. Since the model has been tested on 1:2 matching task, only the V→F task is selected for evaluation.
Verification task: This task is to determine whether the input face and voice belong to the same person or not, and this task can

Table 3: Network classification performance (Acc %)

Task	Unseen-unheard	Seen-heard
Face **G** recognition	99.85	99.80
Voice **G** recognition	97.09	98.75
Face **N** recognition	77.28	85.18
Voice **N** recognition	70.45	80.69

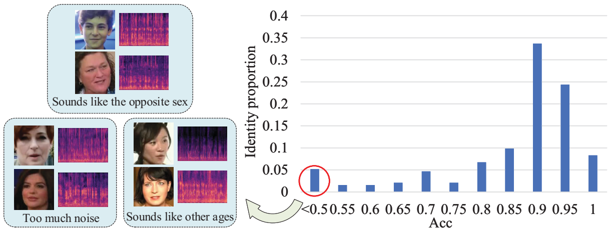

Figure 3: Analysis of low accuracy performance.

be further divided into two seen-heard (tested samples appear in training dataset) and unseen-unheard cases (tested examples not appear in training dataset).

Retrieval task: This task is an extension of cross-modal matching task, and one or more instances might match the given anchor. In this task, 40 face images and 40 voice clips are selected from one person, and mAP metric is selected to validate the cross-modal retrieval performance of the proposed approach.

4.3 Model Performance

Classification results: To validate the model efficiency, we first utilize gender (**G**) and nationality (**N**) to perform classification tasks. Note that, the nationality distribution in the dataset is seriously unbalanced, and it is difficult to conduct efficient training. The classification accuracies obtained by the proposed approach are shown in Table 3, it can be observed that the proposed network modal can well discriminate the meaningful face and voice features to characterize the gender and nationality of the humans. That is, the final outputs of network modal not only can preserve much information about the gender, but also contain valuable information for complex nationality analysis.

1:2 matching results: The recent SVHF-Net [19], LAFV [11], FV-CME [6] and DIMNet [25] are selected for comparison. Similar to these works, the experiments are controlled with varying demographic grouping, particularly '-' denotes the test set are not stratified, '**G**', '**N**' and '**GN**' are stratified by gender, nationality, and grouped gender-nationality, respectively.

As shown in Table 4, it can be found that the proposed ADSM model has yielded the improved 1:2 matching performances in most cases. For V⟶F task, ADSM-TCD improves the DIMNet-IG by 5.13% on unconstrained tests and 9.23% on gender-constrained tests. That is, the mapping operation from original voices and faces to their common covariates is an effective strategy to learn cross-modal embeddings, and the proposed network model provides more capability to learn useful information for cross-modal matching task. As shown in Figure 3, the proposed ADSM network has achieved the state of art in all indicators, the accuracies of most matching results are higher. Meanwhile, we also investigate the low accuracy examples, and there are only few identities whose accuracy is relatively lower. These low accuracies occur on examples with noisy audios and some abnormal associations of faces and voices. Therefore, the adversarial loss generated by modality classifier often contributes to a higher retrieval performance than the model without adversarial learning module.

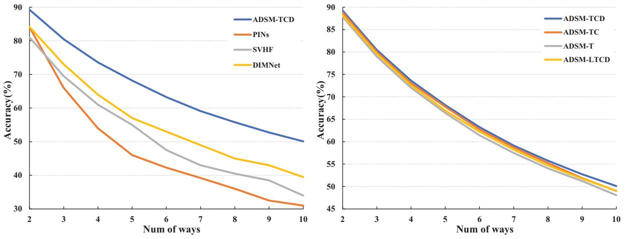

Figure 4: Performances on 1:N matching task.

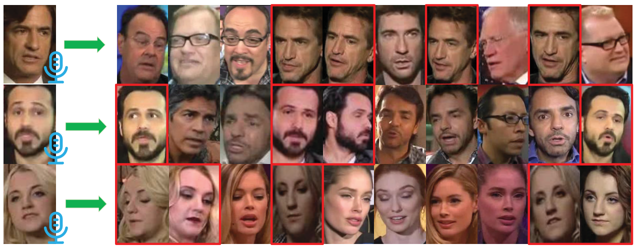

Figure 5: Top 10 cross-modal retrieval results by voice query, and the correctly indexed samples are marked in red.

1:N matching results: For V⟶F, the multi-ways network is tested, which accepts N face samples and predicts the only positive sample from these samples. We also conduct different matching tasks for N>2, as shown in Fig. 4. It can be found that ADSM has achieved the best results on different N values, and training with different data split shows the little effect on the performance. These results are consistent with the results illustrated in Table 4. That is, ADSM is powerful to find the associations between faces and voices.

Verification results: For fair comparison, the data splits are chosen as the same as PINs [18]. Accordingly, the cross-modal verification results in terms of standard AUC value are reported in Table 6, it can be found that the proposed ADSM method outperforms the PINs in different versification tasks, also performs favorably compared to the recent DIMNet work. Therefore, our network model is tolerate to different verification tasks with promising performances.

Cross-modal retrieval results: We also report the cross-modal retrieval results by respectively using voice or face query, whose objective is to retrieve gallery items with the same attribute, *i.e.*, '**G**', '**N**' and '**GN**'. Similar to work [25], 182 identities are randomly selected from the testing set. As shown in Table 4, the proposed ADSM approach always yields the best mAP values, in both V⟶F and F⟶V retrieval tasks. For instance, the mAP score obtained by the proposed ADSM approach reaches up to 64.91% on nationality, which is significantly higher than the result 43.26% obtained by DIMNet-I. Representative retrieval examples are shown in Figure 5, it can be observed that the proposed model is able to well correlate the semantically similar face and voices.

Visualization results: We further utilize the t-SNE [17] algorithm to visualize the learned embedding features, and data examples from eight peoples are randomly selected for visualization. As shown in Figure 6, it can be found that the proposed ADSM network can put the feature embedding of two similar modalities close together, and

Table 4: Cross-modal matching results of voice and face, and the best results are highlighted in bold.

Task	Method	V→F				F→V			
		-	G	N	GN	-	G	N	GN
Matching Accuracy (Acc)	SVHF [19]	81.00	65.20	78.20	-	79.50	-	-	-
	LAFV [11]	78.20	62.90	76.40	61.60	78.60	61.60	76.70	61.20
	FV-CME [6]	78.10	59.97	-	-	77.80	60.65	-	-
	DIMNet-I [25]	83.45	70.91	81.97	69.89	83.52	71.78	82.41	70.90
	DIMNet-IG [25]	84.12	71.32	82.65	70.39	84.03	71.65	82.96	70.78
	ADSM-T	88.27	80.03	83.09	69.13	88.55	79.00	85.82	74.61
	ADSM-TC	88.98	80.38	84.02	69.97	89.15	79.45	86.42	75.36
	ADSM-LTCD	88.77	79.93	83.41	69.09	89.05	79.03	86.02	74.81
	ADSM-TCD	**89.45**	**80.55**	**84.69**	**71.11**	**89.52**	**80.06**	**86.67**	**75.91**
Retrieval Performance (mAP)	FV-CME [6]	1.96	-	-	-	2.18	-	-	-
	DIMNet-I [25]	4.25	89.57	43.26	-	4.17	88.50	43.68	-
	DIMNet-IG [25]	4.42	93.10	43.2	-	4.23	92.16	43.86	-
	ADSM-T	8.97	92.86	64.38	66.28	6.78	91.66	**63.45**	64.36
	ADSM-TC	9.21	**93.34**	**64.91**	**66.46**	7.09	**93.02**	63.42	**65.06**
	ADSM-LTCD	9.16	92.98	59.39	65.83	7.02	92.02	61.75	61.33
	ADSM-TCD	**10.00**	93.18	59.19	61.35	**8.21**	92.12	57.93	59.68

Table 5: The parameter numbers in different network models.

Method	Train set	Test set	Test pairs	Facenet	Facenet params	Voicenet	Voicenet params	All params
SVHF [19]	942	189	0.01M	VGG-M	103.05M	VGG-M	16.62M	123.34M
PINs [18]	901	250	0.03M	VGG-M	103.05M	VGG-M	16.62M	119.67M
FV-CME [6]	862	216	38000M	ResNet-50	23.75M	i-Vector	-	-
DIMNet [25]	924	189	68M	DIMNet-f	8.18M	DIMNet-v	10.43M	18.54M
LAFV [11]	1001	250	0.025M	VGG-16	14.77M	SoundNet	0.86M	15.65M
ADSM-TCD	942	193	221M	ADSM-f	7.02M	ADSM-v	5.38M	**12.40M**
ADSM-LTCD	901	250	374M	ADSM-f	7.02M	ADSM-v	5.38M	**12.40M**

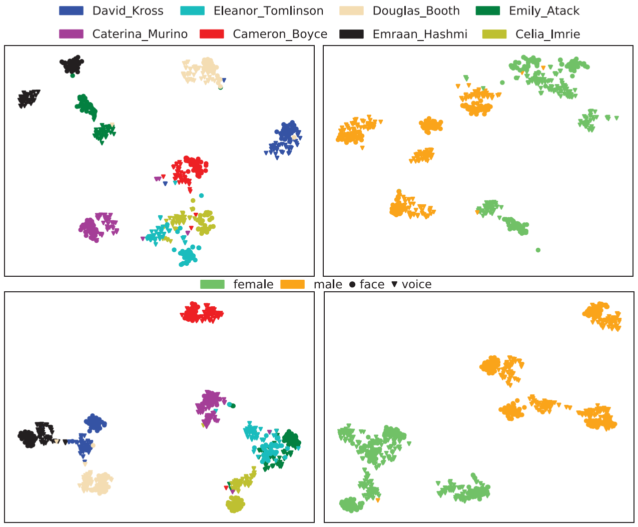

Figure 6: Embedding visualization by identity (left) and gender (right). Top: results by obtained by ADSM-T; Bottom: results learned by ADSM-TCD.

Table 6: AUC values(%) of verification testing.

Task	Method	-	G	N	A	GNA
unseen unheard	PINs [18]	78.5	61.1	77.2	74.9	58.8
	DIMNet [25]	83.2	71.2	81.9	78.0	62.8
	ADSM-LTCD	**88.4**	**79.3**	**83.6**	**83.9**	**64.7**
seen heard	PINs [18]	87.0	74.2	85.9	86.6	74.0
	DIMNet [25]	94.7	89.8	93.2	94.8	87.8
	ADSM-LTCD	**95.9**	**92.5**	**93.9**	**95.5**	**89.8**

the learned feature embeddings are discriminative to benefit the challenging associations between voices and faces.

Ablation Studies and Parameter Analysis: The ablation studies of different learning combinations are shown in Table 4. It can be found that the cross-modal embedding learning module and adversarial learning modules are both beneficial for various voice-face mathching tasks, whereby the proposed ADSM learning framework is capable of capturing the inherent interactions of voice-face more expressively. Besides, deep models often involve large amount of network parameters. Remarkably, the proposed ADSM model is designed to incorporate less model parameters, and the parameter numbers of different network models are elaborated in Table 5. It can be clearly observed that the proposed ADSM network really processes less model parameters, which can greatly reduce the model complexity while holding the learning ability. The experimental results have shown its outstanding performances.

5 CONCLUSION

This paper has presented an efficient adversarial deep semantic matching framework for reliable voice-face interactions and associations. The proposed framework is able to well learn the cross-modal embeddings for voices and faces, while the adversarial learning architecture associated with the proposed multi-modal center loss can well push representations of the same identity closer while pulling those of different identity away. With significantly reduced model parameters, extensive experiments on various cross-modal voice-face matching tasks have verified its outstanding performance.

ACKNOWLEDGMENTS

The work was supported in part by National Science Foundation of China (Nos. 61673185, 61672444, 61972167 and 61976049), in part by Quanzhou City Science & Technology Program of China (No. 2018C107R), in part by State Key Lab. of ISN of Xidian University (No. ISN20-11), in part by Promotion Program for graduate student in Scientific research and innovation ability of Huaqiao University (No. 18013083006), in part by Hong Kong Baptist University, Research Committee, Initiation Grant-Faculty Niche Research Areas (IG-FNRA) 2018/19 (No. RC-FNRA-IG/18-19/SCI/03), in part by ITF of ITC of Hong Kong SAR under Project ITS/339/18, and in part by HKBU Interdisciplinary Research Clusters Matching Scheme with the code: RC-IRCMs/18-19/SCI/01.

REFERENCES

[1] Yewang Chen, Xiaoliang Hu, Wentao Fan, Lianlian Shen, Zheng Zhang, Xin Liu, Jixiang Du, Haibo Li, Yi Chen, and Hailin Li. 2020. Fast density peak clustering for large scale data based on kNN. *Knowledge Based Systems* 187 (2020). no. 104824.

[2] Yewang Chen, Lida Zhou, Songwen Pei, Zhiwen Yu, Yi Chen, Xin Liu, Jixiang Du, and Naixue Xiong. 2019. KNN-BLOCK DBSCAN: Fast Clustering for Large-Scale Data. *IEEE Transactions on Systems, Man, and Cybernetics* (2019). doi:10.1109/TSMC.2019.2956527.

[3] Wentao Fan, Nizar Bouguila, Jixiang Du, and Xin Liu. 2019. Axially Symmetric Data Clustering Through Dirichlet Process Mixture Models of Watson Distributions. *IEEE Transactions on Neural Networks and Learning Systems* 30, 6 (2019), 1683–1694.

[4] Israel D Gebru, Sileye Ba, Georgios Evangelidis, and Radu Horaud. 2015. Tracking the active speaker based on a joint audio-visual observation model. In *CVPR Workshops*. 15–21.

[5] Bashar Awwad Shiekh Hasan, Mitchell Valdes-Sosa, Joachim Gross, and Pascal Belin. 2016. Hearing faces and seeing voices: Amodal coding of person identity in the human brain. *Scientific reports* 6 (2016), 37494.

[6] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. 2018. Face-voice matching using cross-modal embeddings. In *ACM MM*. 1011–1019.

[7] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *CVPR*. 7132–7141.

[8] Frédéric Joassin, Mauro Pesenti, Pierre Maurage, Emilie Verreckt, Raymond Bruyer, and Salvatore Campanella. 2011. Cross-modal interactions between human faces and voices involved in person recognition. *Cortex* 47, 3 (2011), 367–376.

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

[10] Einat Kidron, Yoav Y Schechner, and Michael Elad. 2005. Pixels that sound. In *CVPR*. 88–95.

[11] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. 2018. On learning associations of faces and voices. In *ACCV*. 276–292.

[12] Katharina von Kriegstein, Andreas Kleinschmidt, Philipp Sterzer, and Anne-Lise Giraud. 2005. Interaction of face and voice areas during speaker recognition. *Journal of cognitive neuroscience* 17, 3 (2005), 367–376.

[13] Christoph H Lampert and Oliver Krömer. 2010. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *ECCV*. 566–579.

[14] Xin Liu, Yiu-ming Cheung, Zhikai Hu, Yi He, and Bineng Zhong. 2020. Adversarial Tri-Fusion Hashing Network for Imbalanced Cross-Modal Retrieval. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2020). doi: 10.1109/TETCI.2020.3007143.

[15] Xin Liu, Jiajia Geng, Haibin Ling, and Yiu-ming Cheung. 2019. Attention guided deep audio-face fusion for efficient speaker naming. *Pattern Recognition* 88 (2019), 557–568.

[16] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-ming Cheung. 2019. MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019). doi:10.1109/TPAMI.2019.2940446.

[17] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008), 2579–2605.

[18] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Learnable PINs: Cross-modal embeddings for person identity. In *ECCV*. 71–88.

[19] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*. 8427–8436.

[20] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60 (2020), 101027.

[21] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*. 689–696.

[22] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep face recognition. In *BMVC*. 1:1–1:6.

[23] Nannan Wang, Xinbo Gao, and Jie Li. 2018. Random sampling for fast face sketch synthesis. *Pattern Recognition* 76 (2018), 215–227.

[24] Timothy Wells, Thom Baguley, Mark Sergeant, and Andrew Dunn. 2013. Perceptions of human attractiveness comprising face and voice cues. *Archives of sexual behavior* 42, 5 (2013), 805–811.

[25] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh. 2019. Disjoint mapping network for cross-modal matching of voices and faces. In *ICLR*. https://openreview.net/forum?id=B1exrnCcF7

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*. 1492–1500.

[27] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[28] Xiao-Lei Zhang and Ji Wu. 2012. Deep belief networks based voice activity detection. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 4 (2012), 697–710.

[29] Zhao Zhang, Fanzhang Li, Mingbo Zhao, Li Zhang, and Shuicheng Yan. 2017. Robust Neighborhood Preserving Projection by Nuclear/L2,1-Norm Regularization for Image Feature Extraction. *IEEE Transactions on Image Processing* 26, 4 (2017), 1607–1622.

[30] L Jacob Zweig, Satoru Suzuki, and Marcia Grabowecky. 2015. Learned face–voice pairings facilitate visual search. *Psychonomic bulletin & review* 22, 2 (2015), 429–436.