A BENCHMARK FOR VOICE-FACE CROSS-MODAL MATCHING AND RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal associations between a person's voice and face can be learned algorithmically, and this is a useful functionality in many audio and visual applications. The problem can be defined as two tasks: voice–face matching and retrieval. Recently, this topic has attracted much research attention, but it is still in its early stages of development, and evaluation protocols and test schemes need to be more standardized. Performance metrics for different subtasks are also scarce, and a benchmark for this problem needs to be established. In this paper, a baseline evaluation framework is proposed for voice–face matching and retrieval tasks. Test confidence is analyzed, and a confidence interval for estimated accuracy is proposed. Various state-of-the-art performances with high test confidence are achieved on a series of subtasks using the baseline method (called TriNet) included in this framework. A voice–face dataset of Chinese speakers was constructed to study the generalization ability of language and ethnicity transfer. The source code and dataset will be published along with the paper. The results of this study can provide a basis for future research on using video data for cross-modal learning.

1 INTRODUCTION

Studies in biology and neuroscience have shown that a human's appearance is associated with his or her voice Smith et al. (2016b;a); Mavica & Barenholtz (2013). Both the facial features and voice–controlling organs of individuals are affected by hormones and genetic information Hollien & Moore (1960); Thornhill & Møller (1997); Kamachi et al. (2003); Wells et al. (2013), and human beings have the ability to recognize this association. For example, when speaking on the phone, we can guess the gender and approximate age of the person on the other end of the line. When watching a TV show without sound, we can also imagine the approximate voice of the protagonist by observing his or her face movements. With the recent advances in deep learning, face recognition models Wen et al. (2016); Wu et al. (2018); Liu et al. (2017) and speaker recognition models Wang et al. (2018); Li et al. (2017) have achieved extremely high precision. It is then natural to wonder if the associations between voices and faces could be discovered algorithmically by machines. The research on this problem could benefit many applications such as the synchronization of video faces with talking voices and the generation of faces according to voice.

In recent years, much research attention Wen et al. (2018); Horiguchi et al. (2018); Nagrani et al. (2018a); Kim et al. (2018); Nagrani et al. (2018b) has been paid to voice–face cross-modal learning tasks, which has shown the feasibility of recognizing voice–face associations. This problem is generally formulated as a voice–face matching task and a voice–face retrieval task, as shown in Figure 1. Given a set of voice audio and faces, the aim of voice–face matching is to determine which face has generated a particular voice audio. The aim of voice–face retrieval is to present a ranked sequence of faces according to their relevance with respect to a voice audio query.

The research on this problem is still at an early stage, and a benchmark for this problem still needs to be established. In this paper, we address this issue with the following three contributions: 1) Existing methods are all evaluated on a single dataset of less than 200 identities with limited tasks. Hence, a comprehensive evaluation protocol is created to test the generalization ability of the models. A test confidence metric is proposed for qualifying the statistical significance of experimental results. 2) There are still no publicly available voice–face datasets of Chinese speakers. Hence, dataset of Chinese speakers called CN_VF is constructed, which will facilitate the research of cross-language

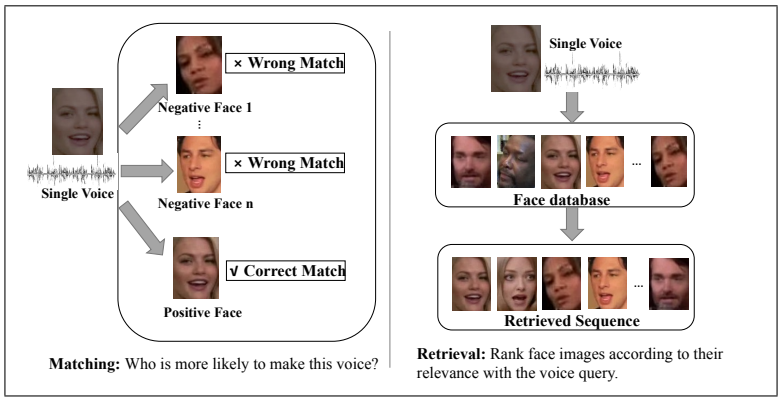


Figure 1: Voice–face cross-modal matching and retrieval tasks.

transfer and be applied in many other scenarios. 3) A solid baseline framework for voice–face matching and retrieval is also proposed. State-of-the-art performances on voice–face matching and retrieval tasks are achieved on large-scale datasets with a high test confidence. Using this framework, the cross-language transfer capabilities of existing models are evaluated on the CN_VF dataset.

2 Tasks, Related Works and Evaluation

2.1 Tasks

In this section, we define the matching and retrieval tasks for voice–face cross-modal learning.

1:2 Matching and 1:n Matching. This is the most common task. Given an audio and two face candidates (only one of which is from the speaker of the audio), the goal is to find the face that belongs to the speaker. The more difficult 1:n matching task is an extension of the 1:2 matching task that increases the number of candidate faces from 2 to N.

Retrieval. Given a "query" voice, the goal of voice–face retrieval is to rank face images according to their relevance with respect to the voice query. This task is a supplement to the matching task. In contrast to 1:n matching, the position-related information of all retrieved faces is also effective for analyzing the model performance. The retrieval task is also significant for many practical applications such as tracing the face of suspect only from the sound file. To date, very few results have been published on this problem, and more research attention needs to be paid to it.

Mean Matching and Retrieval Instead of a single audio segment or single face for one identity, multiple audio segments and faces can provide more information. Matching and retrieval can be conducted on the mean embeddings of multiple audios or images. This is the simplest way to improve the performance of current voice–face matching and retrieval methods. Widespread video resources imply that the use of multiple faces and voices is feasible.

2.2 Related Works

Both the "voice-to-face" and "face-to-voice" matching tasks were studied by Horiguchi Horiguchi et al. (2018) and Pins Nagrani et al. (2018a), who both achieved similar performances on these two tasks. A curriculum learning schedule was introduced in Pins Nagrani et al. (2018a) to mine hard negatives. In Kim's work Kim et al. (2018), various visualizations of embedding vectors were presented to show learned audio–visual associations.

The existing methods for voice–face cross-modal learning can be classified as classification-based methods and pair-wise loss based methods, as shown in Figure 2. CNN-based networks are normally used to embed the voices and faces into feature vectors. SVHF Nagrani et al. (2018b) is a prior study on voice–face cross-modal learning that investigated the performance of a CNN-based deep network on this problem. The human baseline for the voice–face matching task is also presented in this paper. In Wen et al. (2018), a voice or face embedding network is selected to generate common features for

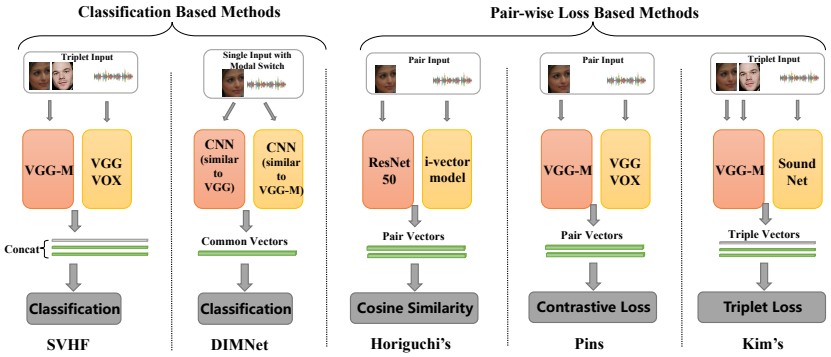


Figure 2: Comparison of existing methods.

either face or voice, and the learning is then supervised by a multi-task classification network and this method obtains an accuracy of 84.12% on 1:2 matching, which exceeds human performance. For pair-wise loss based methods, a pair or a triplet of vectors is embedded by a voice and face network, and contrastive loss Hadsell et al. (2006) or triplet loss Schroff et al. (2015) is used to supervise the learning of the embeddings. The aim of pair-wise loss based methods is to make the embeddings of positive pairs closer and the embeddings of negative pairs farther apart. In contrast, the aim of classification-based methods is to separate the embeddings of different classes. Of these two approaches, pair-wise loss based methods are better at distinguishing hard examples because of the characteristics of this approach. The improved voice-face embedding learning framework proposed in this paper is most similar to Kim's method Kim et al. (2018).

2.3 TEST CONFIDENCE RISK.

The evaluation criteria for matching and retrieval tasks are accuracy and mAPChristopher et al. (2008) respectively. Almost all previous studies Wen et al. (2018); Nagrani et al. (2018a); Kim et al. (2018); Nagrani et al. (2018b) evaluated their methods on a single dataset of less than 200 identities. As shown in the experiment (Section 5.2), the 1:2 matching accuracy tested on multiple datasets with 189 identities varies significantly, from 81% to 87%. ***So the results of all related works that were training and testing on Vox-Vgg-1 are inaccurate.*** Testing a model on a single small dataset may lead to a large deviation in the accuracy. In 1:2 matching task, the accuracy is estimated on the sampled data, to represent the accuracy on the overall population. The estimated accuracy always has a large deviation due to the high sampling risk in the triplet sampling scenario.

Essentially, our aim is to obtain the correct matching probability of a single independent triplet. When the dataset and the model are determined, a single independent sampling conforms to the Bernoulli distribution $B(p)$. The results of n samplings fit the binomial distribution $B(n, p)$. Interval estimation of a binomial distribution can be used for quantifying the deviation of the estimated accuracy. Suppose a dataset D can generate up to N triplets, and the number of sampled triplets used for testing is n. Among the sampled triplets, there are m correctly matched triplets. Suppose the sample rate is p, where $p = \frac{m}{n}$, and the population rate of correctly matched N triplets is P. When n is sufficiently large, p can be approximated as normal distribution $p \sim N(P, \frac{P(1-P)}{n}$. By converting it to a standard normal distribution, we obtain $u = \frac{p-P}{\sqrt{\frac{P(1-P)}{n}}} \sim N(0, 1)$. For a significance level α, the confidence interval of p is $(p - u_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}, p + u_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}})$. Testing a model on multiple datasets is strongly recommended when the dataset is very small. The test can be performed multiple times on datasets with a similar scale, and the results are regarded as conforming to the normal distribution. The t-test can then be used to estimate the confidence interval of the accuracy.

Table 1: Statistics of voice–face cross-modal datasets.Vox-VGG-n represents the combined dataset of the VoxCeleb Nagrani et al. (2017); Chung et al. (2018) and VGGFace Cao et al. (2018); **?**); Parkhi et al. (2015) and n means the version. The number of images refers to the number of all images remaining after MTCNN Zhang et al. (2016) face detection.

Dataset	Gender	#Identities	#Utterances	#Images
Vox-VGG-2	Male	3,682	775,260	1,268,937
	Female	2,312	316,749	636,079
	Total	5,994	1,092,009	1,905,016
Vox-VGG-1	Male	690	103,295	378,152
	Female	561	50,221	195,131
	Total	1,251	153,516	573,283
CN_VF	Male	464	960,764	233,745
	Female	534	1,259,315	271,904
	Total	998	2,220,079	505,649

3 DATASETS

Currently available voice–face datasets are the data generated by the common speakers of VGGFace Simonyan & Zisserman (2014) face recognition dataset and VoxCeleb Nagrani et al. (2017) speaker recognition dataset. The voice–face datasets have two versions, Vox-VGG-1 Cao et al. (2018); Simonyan & Zisserman (2014) and Vox-VGG-2 Chung et al. (2018); Nagrani et al. (2017), which include 1,251 and 5,994 identities, respectively. Almost all previous research evaluate their methods on Vox-VGG-1. In this paper, both Vox-VGG-2 and Vox-VGG-1 are used to evaluate the proposed baseline method, TriNet.

A new dataset of Chinese speakers called CN_VF is presented in this paper. Collecting datasets from speakers of different languages is important for analyzing the characteristics of cross-language cross-modal learning, and publicly available voice–face datasets of Chinese speakers are still unavailable. The dataset was constructed by extracting face and audio data from sources such as MOOCs, lectures, vlogs, and talk shows. As presented in Table 1, the overall scale of CN_VF is similar to that of Vox-VGG-1.

4 TRINET BASELINE METHOD

4.1 BASELINE METHOD FOR VOICE–FACE MATCHING AND RETRIEVAL: TRINET

The proposed baseline framework is shown in Figure 3.

4.1.1 TRIPLET MINING.

The sampled triplets need to be mined from the datasets. There are two ways to mine triplets, namely online mining and offline mining Yu et al. (2018). In online mining, a batch of identities is randomly selected first, and then a certain number of face images and audio segments for each identity of the batch are sampled. For example, suppose four identities are sampled, and four faces and four voices are sampled for each identity. As a result, a total of 3,072 triplets are used as input. In offline mining, discrete triplets are randomly mined to create a single input each time. In this study, online mining is adopted for training and testing efficiency.

4.1.2 FEATURE EXTRACTION NETWORK.

LResNet50 He et al. (2015) and Thin ResNet34 Xie et al. (2019) with NetVLAD Arandjelovic et al. (2017) are networks that perform well on face recognition and speaker recognition tasks, respectively. For a specific triplet $< v^{(i)}, f^{(i)}, f^{(j)} >$, $v^{(i)}$, and $f^{(i)}$ are from the same identity, and $v^{(i)}$ and $f^{(j)}$ are from different identities. The feature extraction functions for voice and face are defined as $Feature_v(v)$ and $Feature_f(f)$, respectively, and a fully connected layer is added to form the embedding vectors as $emb_v(v) = s \times \|W_v \times Feature_v(v) + B_v\|_2^2$ and $emb_f(f) = s \times \|W_f \times Feature_f(f) + B_f\|_2^2$. Finally, we add L2 constraints and scale variables to the output of the network.

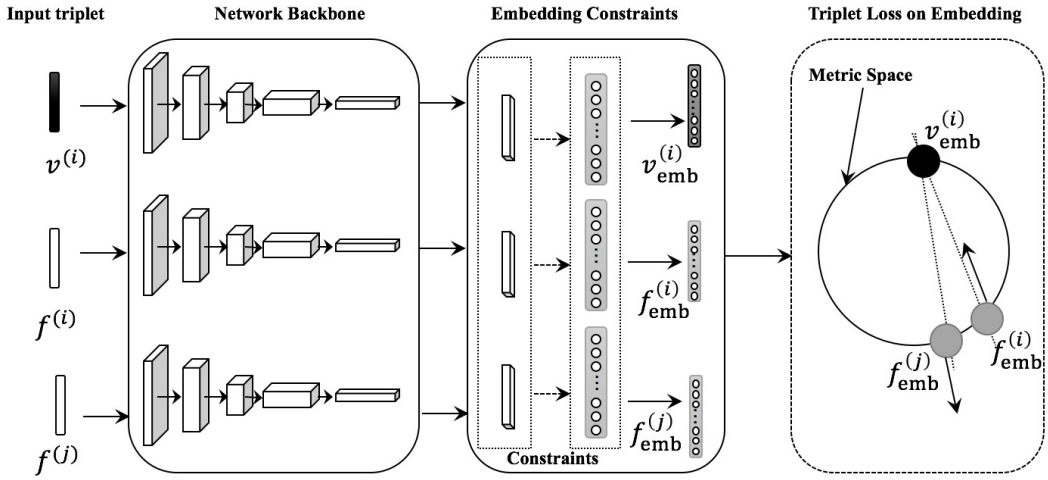


Figure 3: Overview of TriNet. 1) A feature extraction module extracts voice and face features. 2) Embeddings are constrained to lie on a spherical space using L2 normalization. In this step, the audio network is frozen. 3) The triplet loss function is used to learn the embeddings.

4.1.3 Embedding Constraints and the Loss Function.

As illustrated in Figure 4a, embedded vectors from the same person will appear in a Euclidean space after a long period of training. Because there are billions of input triples, it is difficult to obtain satisfactory results by directly training in a huge Euclidean space. To deal with this problem, two strategies are adopted in the method proposed in this paper. First, L2 normalization is added to constrain the embedding vectors to a spherical space (Figure 4b). Second, voice-anchored embedding learning is adopted. By freezing the pre-trained voice embedding network, feature vectors from voice serve as anchors, and the goal of the model is to make positive instances approach each other while keeping the negative instances away (Figure 4c). Examples tend to be distinguished much better and faster in the voice-anchored embedding learning process when used with the L2 constrained space. Triplet loss is adopted in this paper for embedding learning. Suppose $d(x, y)$ indicates Euclidean distance; the loss function is defined as

$$Loss = \sum_{v^{(i)}, f^{(i)}, f^{(j)}, i \neq j} \max\{d(emb_v(v^{(i)}), emb_f(f^{(i)})) - d(emb_v(v^{(i)}), emb_f(f^{(j)})) + m, 0\}, \tag{1}$$

where m is a margin to control the distance between positive and negative pairs.

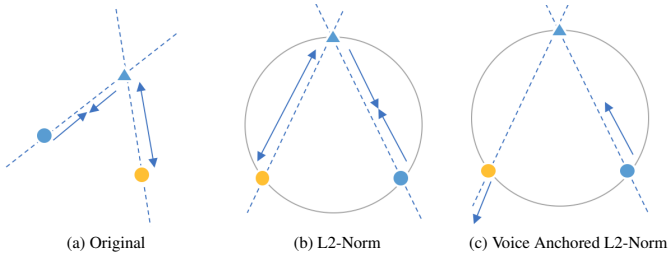


Figure 4: Visualization of three metric learning methods.

5 EXPERIMENT

5.1 BASELINE MODEL SETUP

Training Details. TriNet was trained on Vox-VGG-2 with 5,994 identities. Face detection based on MTCNN Zhang et al. (2016) was conducted and all face images were then rescaled to $112 \times 112 \times 3$ to form the input for the face embedding networks. Audio preprocessing consisted of 512-point fast Fourier transform, a short-time Fourier transform Benesty et al. (2011) for each frame, and normalization. The audio segments used for training were uniformly trimmed to 2.5 s for training efficiency, and no the test audio segments were not clipped. The input shape for a k-s audio clip is $257 \times (100 \times k) \times 1$. The voice embedding network and face embedding network were pre-trained by VoxCeleb2 and VGGFace2, respectively. Margin, m, for the triplet loss was set to 1, and scale, s, for the L2 normalization was set to 128. The Adam optimizer was adopted in these experiments, and the total number of learning steps was 70k. The learning rates of the final fully connected layer for $step < 20k$, $20k < step < 40k$, $40k < step < 60k$, and $step > 60k$ were $10^{-3}, 10^{-4}, 10^{-5}$, and 10^{-6} respectively. The learning rate of the face embedding network was fixed to 10^{-6}.

Testing Details. 1) For the 1:2 matching task, a total of 10,000 steps were tested on the baseline, which implies that a total of 30.72 million triples were tested. Note that the gender of the test triples in the 1:2 matching task was balanced. 2) For the 1:n matching task, the number of tuples to be sampled will be much higher than the number of triples in 1:2 matching; therefore, we performed this test directly on the 10k tuples. This will lower the confidence level, but the results are still useful for comparisons. 3) For the retrieval task, a face database of 500 pictures was constructed from 100 randomly selected identities, and 40 audio queries were constructed for each identity.

5.2 PERFORMANCE AND COMPARISON

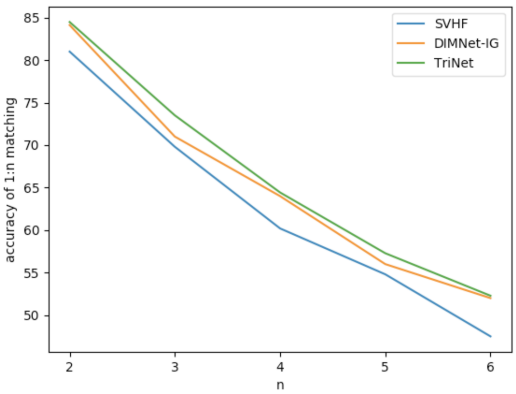
Table 2: Comparison of the proposed method with other models on the 1:2 matching task and retrieval task."ACC" is accuracy.

Tasks	**Method**	**Test Identities**	**Test Triplets**	**ACC(%)**
1:2 Matching	TriNet	1,251	30M	$84.48 \pm 0.01\%$
	SVHFNagrani et al. (2018b)	189	10k	$81.00 \pm 1.01\%$
	DIMNet-IGWen et al. (2018)	189	678M	84.12%
	Kim's Kim et al. (2018)	250	-	78.20%
	Horiguchi's Horiguchi et al. (2018)	216	38B	77.80%
Retrieval	**Method**	**Test Identities**	**Chance (mAP%)**	**mAP(%)**
	TriNet	1251	2.15	11.48
	DIMNet-IG Wen et al. (2018)	189	1.07	4.42
	Horiguchi's Horiguchi et al. (2018)	216	0.46	1.96

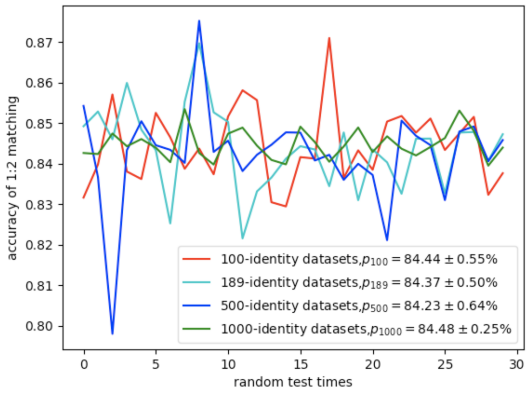
5.2.1 COMPARISONS ON MATCHING AND RETRIEVAL TASKS.

Comparisons of TriNet and related work on the 1:2 matching task and retrieval task are shown in Table 2. TriNet achieves state-of-the-art performance on these two main tasks. As shown in Figure 5b, on the 1:n task, the performances of all the methods decrease rapidly as n increases. This task is still significantly difficult. Some TriNet retrieval results that fit $p@1 = 1$ are illustrated in Figure 5a. The top ranked faces in each sequence are very similar to the target face. The results of 1:2 matching on the mean voice and mean face are shown in Table 3. Two variables, m_f and m_v, are introduced to represent the number of faces and audios used to compute the mean embedding. Various values of m_f and m_v are tested for 1:2 matching. For retrieval, m_v was set to 20 and m_f was set to 5. As shown in the table, this simple strategy of using multiple faces and voices can further improve the matching accuracy and retrieval mAP. Specifically, when $m_v = m_f = 10$, an accuracy of $89:66 \pm 0:80\%$ can be obtained for TriNet on the 1:2 matching task, which is 5% higher than that of single voice and single face matching. This improvement reveals a broad prospect for future research on using video data for cross-modal learning.

(a) Qualitative analysis of retrieval results produced by TriNet.



(b) Comparisons with other available results on the 1:n matching task.



(c) Changes in accuracy for experiments on four different sizes of dataset.

Figure 5: (a) Retrieval results. (b) Comparisons on the 1:n matching task. (c) Accuracy curves.

Table 3: Performance of 1:2 matching and retrieval on mean face and voice embeddings. Here, m_f and m_v denote the number of face images and audio segments, respectively, used to calculate the averages. The values of m_v and m_f are 1 by default.

Matching on Mean Voices and Mean Faces (TriNet)					
m_f	ACC(%)	m_v	ACC(%)	$m_f + m_v$	ACC(%)
5	85.55	5	85.13	1+1	84.48
10	84.56	10	86.01	5+5	86.42
15	86.28	15	85.49	10+10	**89.66**
20	86.53	20	86.16	20+20	89.54
30	84.71	30	85.63	30+30	89.28

Retrieval on Mean Voice and Mean Face		
	mean voice mAP(%)	**mean face mAP(%)**
TriNet	12.53	21.65
Random	2.15	5.25

Table 4: Language and ethnicity transfer experiments on the CN_VF dataset. "TriNet-CN:" denotes the default TriNet, which is tested on datasets of Chinese speakers; "-ft" denotes that the model is fine-tuned on datasets of Chinese speakers; "DIMNet (reproduce)" denotes the reproduced DIMNet-IG model trained on Vox-VGG2; and "SVHF (reproduce)" denotes the reproduced SVHF model trained on Vox-VGG2.

Model	Training	Test	1:2(%)	1:3(%)	Male(%)	Female(%)	mAP(%)
TriNet	5994 (EN)	1251 (EN)	84.48	73.50	69.85	71.04	11.48
TriNet-CN	5994 (EN)	199 (CN)	71.51	55.38	55.57	54.46	5.00
DIMNet (reproduce)	5994 (EN)	199 (CN)	70.30	53.96	52.02	52.71	4.74
SVHF (reproduce)	5994 (EN)	199 (CN)	69.02	-	51.11	50.45	-
TriNet-CN-ft	5994+799 (EN+CN)	199 (CN)	**79.43**	**67.10**	**65.10**	**63.60**	**8.55**
DIMNet-ft (reproduce)	5994+799 (EN+CN)	199 (CN)	75.15	58.18	55.94	59.64	7.20
SVHF-ft (reproduce)	5994+799 (EN+CN)	199 (CN)	74.95	-	52.26	56.48	-

5.2.2 Test Confidence of Datasets with Different Scales.

Figure 5c shows the fluctuations in the estimated accuracies of TriNet on the 1:2 matching task when 30 repeated random tests were conducted. The numbers of sampled identities for each curve are 100, 189, 500, and 1,000. For a determined dataset scale (such as 100 identities), instead of testing the model on a single dataset with 100 identities, 30 randomly sampled sets of data with 100 identities were used for testing. As shown in the figure, when a small-scale dataset is used, the accuracy of different runs fluctuates substantially. For large datasets, there are fluctuations in test accuracies; however, in general, the test results are more generalized; therefore, large datasets are strongly recommended for evaluation.

5.3 Language and Ethnicity Transfer Ability Evaluation

The language and ethnicity transfer abilities of models were also analyzed. Two well-performing existing models were reproduced and trained with Vox-VGG-2 for comparison. As shown in Table 4, when we directly transfer the models trained on datasets of English speakers to datasets of Chinese speakers, the performances of the models substantially decreases. When the models are fine-tuned on the datasets of Chinese speakers, the performance is significantly improved. Specifically, the 1:2 matching accuracy of TriNet on the datasets of Chinese speakers is improved by approximately 8% after fine tuning. It is believed that better results will be achieved if much larger datasets used.

6 Conclusion

In this study, a benchmark was established for voice–face matching and retrieval. The contributions of this paper are as follows:

A solid voice–face matching and retrieval baseline method (TriNet) was proposed, which was tested on large-scale dataset with comprehensive ablation studies. The test confidence was proposed as a metric for qualifying the statistical significance of the experiments. On the 1:2 matching and retrieval tasks, TriNet achieved an accuracy of 84.48% and a mAP of 11%. Compared with the best results published so far, there is a 7% improvement in mAP. Using mean face and mean voice embeddings, the matching accuracy and retrieval mAP can be further improved by approximately 5% and 10%, respectively. This improvement implies a broad prospect for future research on using video data for cross-modal learning.

A voice–face dataset of 998 Chinese speakers was constructed to study the language transfer ability of the models. To the best of our knowledge, this is the first attempt to study the cross-language transfer problem of voice–face matching in the literature. By fine tuning the pre-trained model on the 1:2 matching task, an accuracy of 79% was obtained for Chinese speakers. Much better results can be achieved on further expanding the scale of the data.

REFERENCES

R Arandjelovic, P Gronat, A Torii, T Pajdla, and J Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP (99):1–1, 2017.

Jacob Benesty, Jingdong Chen, and Emanuël AP Habets. *Speech enhancement in the STFT domain*. Springer Science & Business Media, 2011.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

D Manning Christopher, Raghavan Prabhakar, and Schütze Hinrich. Introduction to information retrieval. *An Introduction To Information Retrieval*, 151(177):5, 2008.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. 2018.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015.

Harry Hollien and G. Paul Moore. Measurements of the vocal folds during changes in pitch. *Journal of Speech and Hearing Research*, 3(2):157–165, 1960.

Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. Face-voice matching using cross-modal embeddings. In *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 1011–1019. ACM, 2018.

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.

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*, pp. 276–292. Springer, 2018.

Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

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.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. 2017.

Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 71–88, 2018a.

Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8427–8436, 2018b.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, pp. 6, 2015.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

Harriet MJ Smith, Andrew K Dunn, Thom Baguley, and Paula C Stacey. Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, 14(1): 1474704916630317, 2016a.

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, 2016b.

Randy Thornhill and Anders Pape Møller. Developmental stability, disease and medicine. *Biological Reviews*, 72(4):497–548, 1997.

Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. Speaker diarization with lstm. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5239–5243. IEEE, 2018.

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

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh. Disjoint mapping network for cross-modal matching of voices and faces. 2018.

Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5791–5795. IEEE, 2019.

Baosheng Yu, Tongliang Liu, Mingming Gong, Changxing Ding, and Dacheng Tao. Correcting the triplet selection bias for triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 71–87, 2018.

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

A Appendix

You may include other additional sections here.