Cross-Modal Distillation for Speaker Recognition

Yufeng Jin[1], Guosheng Hu[2], Haonan Chen[3], Duoqian Miao[1], Liang Hu[1], Cairong Zhao[1]*

[1] School of Electronic and Information Engineering, Tongji University, China
[2] Oosto, UK
[3] Alibaba Group, China
jinyufeng@tongji.edu.cn, huguosheng100@gmail.com, haolan.chn@alibaba-inc.com, dqmiao@tongji.edu.cn, rainmilk@gmail.com, zhaocairong@tongji.edu.cn

Abstract

Speaker recognition achieved great progress recently, however, it is not easy or efficient to further improve its performance via traditional solutions: collecting more data and designing new neural networks. Aiming at the fundamental challenge of speech data, i.e. low information density, multimodal learning can mitigate this challenge by introducing richer and more discriminative information as input for identity recognition. Specifically, since the face image is more discriminative than the speech for identity recognition, we conduct multimodal learning by introducing a face recognition model (teacher) to transfer discriminative knowledge to a speaker recognition model (student) during training. However, this knowledge transfer via distillation is not trivial because the big domain gap between face and speech can easily lead to overfitting. In this work, we introduce a multimodal learning framework, VGSR (Vision-Guided Speaker Recognition). Specifically, we propose a MKD (Margin-based Knowledge Distillation) strategy for cross-modality distillation by introducing a loose constrain to align the teacher and student, greatly reducing overfitting. Our MKD strategy can easily adapt to various existing knowledge distillation methods. In addition, we propose a QAW (Quality-based Adaptive Weights) module to weight input samples via quantified data quality, leading to a robust model training. Experimental results on the VoxCeleb1 and CN-Celeb datasets show our proposed strategies can effectively improve the accuracy of speaker recognition by a margin of $10\% \sim 15\%$, and our methods are very robust to different noises.

Introduction

A wealth of information is contained in the speaker's voice which can be abstracted into different properties, such as gender, age, tone, etc. These properties can contribute to identity recognition, i.e. speaker recognition. Speaker recognition is widely applied in the real world. For example, in some smart audio devices, personalized configurations can be loaded by recognizing the speaker. In addition, it has important applications in security systems, investigation, forensics, etc. I-vector (Dehak et al. 2010) is a traditional speaker recognition method. With the popularity of deep learning, neural networks (Nagrani, Chung, and Zisserman 2017; Snyder et al. 2018; Okabe, Koshinaka, and Shinoda 2018) started to become the mainstream and achieved promising progress. However, speaker recognition is *fundamentally* difficult, the voices of two speakers are likely very similar, since the population (the number of speakers) can be very large and the information density of speech is very low. The intuitive solution is to collect bigger data and design more effective neural networks, however, this solution maybe not very efficient and does not approach the fundamental difficulty: the low information density of speech data.

Multimodal learning mitigates the challenges of speaker recognition at their root by introducing richer and more discriminative information as input for speaker recognition. For identity recognition, face images clearly convey much more identity-related information than voice, and faces are more widely applied than speech for identity recognition. It inspires researchers to use both face and speech for identity recognition (Tao, Das, and Li 2020; Sarı et al. 2021; Qian, Chen, and Wang 2021). Their experimental results show that the multimodal speaker recognition is much stronger than that using speech only as input. However, their multimodal speaker recognition assumes face and speech are always available during the training and test periods. Actually, many applications have only speech input and do not have face images as input, e.g. most smart speakers do not have cameras.

Multimodal learning motivates us to ask a question: can we use multimodal learning for the scenarios of speech input only? This question triggers us to bridge the multimodal and speech only identity recognition by introducing a new setting: training data with both annotated speech and face data and test set with speech data only. Existing works (Inoue 2021; Cai, Wang, and Li 2022) have a similar multimodal setting, but their training data is unlabelled.

We think our aforementioned setting is technically feasible. This feasibility results from (1) technical soundness and (2) available data. For (1), we assume the two modalities (face and speech) are highly correlated for identity recognition. In this way, the feature space of the networks trained separately by these two modalities is also correlated. Based on this assumption, the stronger modality (face) can transfer the knowledge in feature space to the weaker modality

*Corresponding author

(speech) during the training, making the student network more discriminative than that without teacher (face model) supervision. During the test period, the speech network only, which already learns the discriminative knowledge from the teacher, can conduct robust identity recognition. This process is illustrated in Fig. 1. Then we explain the soundness of the assumption where face and speech are very correlated for identity recognition. Biological studies (Kamachi et al. 2003; Smith et al. 2016) show that human speech is correlated with facial appearance, and some attributes (such as gender, age, race, and some hormone levels) influence both appearance and voice. The success of voice-face cross-modality matching (Nagrani, Albanie, and Zisserman 2018) and speech-to-face generation (Duarte et al. 2019; Oh et al. 2019) tasks also demonstrates that this correlation can be learned by neural networks. For (2), fortunately, a multimodal recognition dataset (Chung, Nagrani, and Zisserman 2018) with aligned face and speech data is available. With this dataset, we can investigate our solutions under the aforementioned setting.

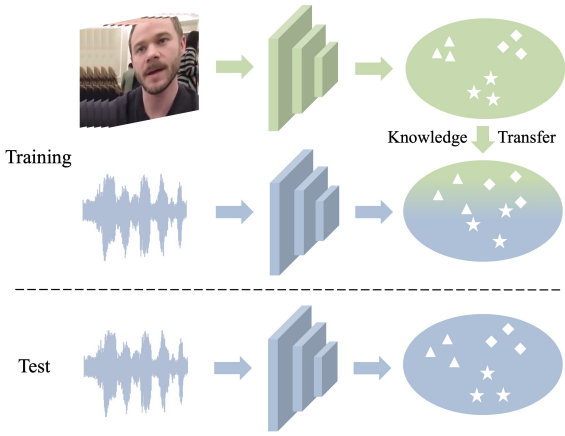

Figure 1: Cross-modal knowledge transfer. The face model (teacher) transfers the discriminative knowledge to the speech model (student) during training. During test, the speech model only conducts identity recognition. The same shape (star, triangle, diamond) represents the same identity.

To achieve the knowledge transfer from a stronger modality (face) to a weaker one (speech) during training, we introduce knowledge distillation. Knowledge distillation is a technique to supervise the training of a student network by a stronger teacher model. Knowledge distillation is well studied and has widely been applied to speech analysis, computer vision, etc. From the source of knowledge, knowledge distillation can be categorized as feature-based, relation-based, and response-based methods (Gou et al. 2021). Feature-based methods (Romero et al. 2015) conduct distillation using the output from the last or intermediate layers of the neural network. Relation-based methods (Tung and Mori 2019; Park et al. 2019; Peng et al. 2019) capture the relationship between different samples or different layers as knowledge. Response-based methods (Hinton et al. 2015) use the output of the last layer of networks, i.e. logits for knowledge transfer.

However, we empirically find that simply applying existing knowledge distillation methods to our cross-modal teacher-student learning cannot achieve desirable performance. We think two possible reasons lead to this degraded performance: (1) the big domain gap in latent space between two modalities and (2) the quality of input data (face and speech) not well aligned. Specifically, for (1), most existing knowledge distillation methods minimize the difference ($l1$ or $l2$ norm) between the teacher and student, forcing the student to behave exactly the same as the teacher. However, the domain gap between two modalities (face and speech) is clearly big and forcing them to be exactly the same can easily lead to overfitting. For (2), empirically, if the data quality of two streams (face and speech) does not match, e.g. a very blurry face supervises an audio sequence of good quality, the performance of distillation will be degraded.

In this paper, we propose a VGSR (Vision-Guided Speaker Recognition) method which can improve the accuracy and generalizability of speaker recognition. Specifically, we propose a distillation strategy, MKD (Margin-based Knowledge Distillation), which introduces a *loose* constraint with a margin between two modalities instead of forcing them to be exactly the same. Our MKD can facilitate the student to learn the discriminative features while avoiding overfitting to learn irrelevant features. Furthermore, our MKD can easily adapt to many mainstream distillation loss functions. Specifically, we reformulate the existing feature-based distillation (Romero et al. 2015), relation-based distillation (Tung and Mori 2019), and response-based distillation (Hinton et al. 2015) loss functions to adapt them to our cross-modality knowledge transfer. As aforementioned, the quality of input data can greatly affect the distillation performance. Thus, we propose a weighting method, QAW (Quality-based Adaptive Weights), which quantifies the data quality by $l2$ norm and then weights the samples using the quality scores for distillation, greatly improve the model robustness.

Our method is trained on the multimodal dataset VoxCeleb2 (Chung, Nagrani, and Zisserman 2018) and evaluated on the speech dataset in VoxCeleb1 (Nagrani, Chung, and Zisserman 2017) and CN-Celeb (Fan et al. 2020). The results show that the VGSR method can effectively improve the performance of speaker recognition (around $10\% \sim 15\%$). In addition, since our method is quality-aware, and less affected by low quality samples, showing promising model robustness, e.g. robustness against noises.

Our contributions can be summarized as:

- We introduce a practical and technically feasible setting to the society: annotated face and speech data for training, and speech data only for test. Based on this setting, we propose a cross-modality learning method VGSR, leading to the promising accuracy and model robustness for speaker recognition.
- Due to the big domain gap of different modalities, the existing knowledge distillation methods do not work well for cross-modality knowledge transfer. We propose a distillation strategy, MKD, which introduces a *loose* align-

ment between the teacher and student, to effectively avoid overfitting caused by the modality domain gap. Our MKD can work in a plug-and-play way, easily adapting to existing distillation loss functions.

- We propose a quality-aware sample weighting method, QAW, which can improve the robustness of our method, effectively avoiding the negative effect caused by low quality samples.
- We conduct extensive experiments on VoxCeleb1, VoxCeleb2, and CN-Celeb, the results show that our method can effectively improve the accuracy of speaker recognition, and the robustness against noises.

Related Work

Speaker Recognition

Speaker recognition is the task of voice-based biometric identification, which plays an important role in smart voice assistants. I-vector (Dehak et al. 2010) is a traditional speaker recognition method.

Recently, the deep neural network has achieved better performance in speaker recognition than the traditional method. Snyder et al. (Snyder et al. 2018) proposed X-Vector, an early speaker recognition model for deep learning. Chung et al. (Chung, Nagrani, and Zisserman 2018) propose the VoxCeleb2 dataset, which is one of the most commonly used speaker recognition datasets, and establish a new benchmark using the ResNet model (He et al. 2016). Okabe et al. (Okabe, Koshinaka, and Shinoda 2018) propose the attentive statistics pooling method, which does attention weighting and adds statistics when pooling. Chung et al. (Chung et al. 2020) compare various loss functions in speaker recognition tasks and proposes angular prototypical loss to achieve the best results. Desplanques et al. (Desplanques, Thienpondt, and Demuynck 2020) propose a powerful speaker recognition model ECAPA-TDNN.

Some researchers also try to combine face images and audio for audio-visual speaker recognition. With the addition of visual modalities, promising recognition performance can be achieved. Sar et al. (Sarı et al. 2021) propose to learn joint audio-visual embeddings and perform cross-modal verification. Qian et al. (Qian, Chen, and Wang 2021) proposes and compares multiple multimodal fusion methods. However, this type of method may encounter the problem that the face image cannot be acquired when it is used in practice. Unlike them, we use speech input only during test.

Knowledge Distillation

Knowledge distillation is a technique that uses a high-performance teacher model to guide student model training and is often used for model compression. As for the type of knowledge, Gou et al. (Gou et al. 2021) classifies knowledge distillation into three categories: feature-based, relation-based, and response-based. Feature-based methods use features from the last or intermediate layers to distill. Zagoruyko and Komodakis (Zagoruyko and Komodakis 2017) propose the attention transfer method, which encouraging the student model to learn the spatial attention distribution of the teacher model. Relation-based methods use relationships between samples. Tung and Mori (Tung and Mori 2019) propose the similarity preserving method, encouraging the similarity of activations between samples and samples of the teacher model and the student model to be consistent. Response-based methods use the output of the last layer of the model, the logits, for distillation. The earliest knowledge distillation method (Hinton et al. 2015) is achieved by minimizing the kl divergence between the logits of the teacher model and the student model.

As for cross-modal knowledge distillation, most methods operate between two very similar modalities. Gupta et al. (Gupta, Hoffman, and Malik 2016) propose a method to use a model of RGB images as a teacher model to guide the training of depth and optical flow image models. Tian et al. (Tian, Krishnan, and Isola 2019) proposes contrastive representation distillation, which implements cross-modal distillation from RGB images to depth images. Some methods that distill between disparate modalities usually use the teacher modality to provide labels for the student model in an unsupervised setting. Inoue (Inoue 2021) uses the face recognition model to provide positive and negative pairs for unlabeled speech data, and uses metric learning to train the speaker recognition model. Unlike them which use an unsupervised setting, in this work, we explore a supervised cross-modal distillation. Zhang et al. (Zhang, Chen, and Qian 2021) use a multi-modal teacher of face and speech to guide the training of single-modal student, and find that the gap between the speech and the teacher system is large, making it difficult to improve the performance. This paper focuses on solving problems such as large modal gaps.

Method

Our method uses a dual stream of visual and speech inputs in the training phase to transfer knowledge from the visual modality to the speech modality. In the test phase, the speech model only conducts identity recognition.

Algorithmic Overview

The overall structure of our proposed method VGSR (Vision-Guided Speaker Recognition) is shown in Fig. 2. We have two networks, teacher (face encoder) and student (audio encoder). The teacher network transfers the knowledge to student via the proposed Margin-based Knowledge Distillation (MKD). To make the MKD more robust, the Quality-based Adaptive Weight (QAW) is used to weight the samples based on data quality to avoid the negative effects of low-quality data. In practice, we find the performance is not satisfying if we directly use face features from a pre-trained face recognition model because of the big domain gap between face and speech. We use a projection head, a three-layer MLP, to narrow down the domain gap.

Formally, we have a face image I_T, subscript $_T$ means teacher. The image is encoded by a pre-trained face encoder $Encoder_f$ and we obtain the features E_T as shown in Eq. (1). In order to extract speech-related features, the original features E_T are further fine-tuned through the projection head, a three-layer MLP. To avoid overfitting by fine-tuning, the original features E_T and the fine-tuned features

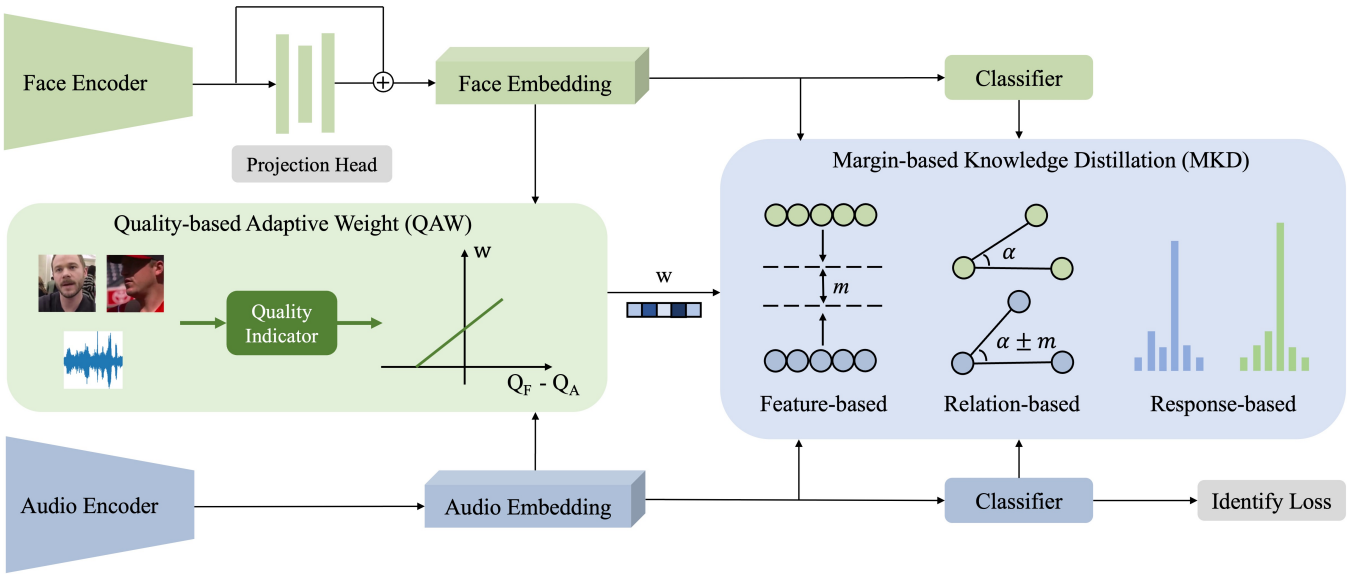

Figure 2: The overview of VGSR method where the teacher (face recognition model) transfers discriminative knowledge to the student (speech model). The MKD strategy enables effective cross-modal knowledge transfer, greatly reducing overfitting. MKD can work with diverse distillation strategies, i.e., feature-, relation- and response-based. QAW module can weight the training samples based on the quantified data quality, leading to a robust training. A pretrained face recognition model is used and it is fine-tuned by a projection head.

are mixed in a certain ratio α, to obtain the final teacher features F_T, as shown in Eq. (2).

$$E_T = Encoder_f(I_T) \tag{1}$$

$$F_T = \alpha \cdot E_T + (1-\alpha) \cdot MLP(E_T) \tag{2}$$

As for the speech branch, the input speech is encoded by the audio encoder to obtain the student feature F_S. The total loss is the combination of the identity loss supervised with label information in speaker recognition and the distillation loss MKD, and the distillation loss has an adaptive sample weight $\mathbf{w}$ generated by the QAW module,

$$\mathcal{L} = \mathcal{L}_{identity}(F_S, label) + \mathbf{w} \cdot \mathcal{L}_{distillation}(F_S, F_T) \tag{3}$$

We detail MKD and QAW in the next section.

Margin-Based Knowledge Distillation (MKD)

Empirically, we cannot achieve desirable performance if we simply use the existing knowledge distillation method. After extensive experiments and analysis, we realize the existing knowledge distillation methods use $l1$ or $l2$ loss to force the student to learn to be exact to the teacher. It works well if the domain gap between teacher and student is relatively small. However, in our task, the cross-modality domain gap between face and speech is very big. The existing knowledge distillation methods tend to cause overfitting. Motivated by metric learning (Schroff, Kalenichenko, and Philbin 2015) which usually uses a margin to separate positive and negative pairs, we introduce a margin m to relatively loosely align the teacher and student. In this way, we do not push the student forward to be exactly the same as the teacher. Instead, we use a margin m to bound the maximal similarity between face and speech. This margin brings a mechanism that can potentially ask the student to learn the discriminative information from the teacher and effectively avoid overfitting.

The existing distillation methods can simply categorized as feature-based, relation-based, and response-based according to Gou et al. (Gou et al. 2021). Our Margin-based knowledge distillation (MKD) can easily adapted to these mainstream distillation methods. Then we formulate MKD for different distillation methods.

Feature-based. Feature-based knowledge distillation aligns the teacher and student using the output from the last or intermediate layers of the model. We empirically find the cosine similarity works better than $l2$ similarity for feature-based distillation. Cosine similarity requires angular similarity rather than numerical equality.

Our MKD introduces a hyperparameter m to construct the distillation loss over cosine similarity,

$$\mathcal{L}_{fea} = \left\lfloor m - \frac{F_T \cdot F_S}{\|F_T\| \cdot \|F_S\|} \right\rfloor \tag{4}$$

where $\lfloor\rfloor$ means cut down to 0, equivalent to $max(\cdot, 0)$. Clearly, when the similarity between the two modalities reaches m, the loss becomes 0.

Relation-based. Relation-based knowledge distillation usually exploits the relationship between different samples. The typical method is to maintain the similarity between samples, which helps to learn the structural features in the teacher model. Specifically, after obtaining the output feature $F \in \mathbf{R}^{b \times c}$, where b is the batch size and c is the num-

ber of channels of the feature map, we calculate the cosine similarity between every two samples, and a $b \times b$ similarity matrix can be obtained, as shown in Eq. (5).

$$F_n = F / \|F\|; \quad G = F_n \cdot F_n^T \tag{5}$$

Then the goal of the distillation loss is to let the student model learn the similarity of the teacher model. Our MKD introduces a hyperparameter m to achieve,

$$\mathcal{L}_{rel} = \left\lfloor (G_T - G_S)^2 - m \right\rfloor \tag{6}$$

Clearly, the loss becomes 0 when the difference of two similarity matrices from two modalities is smaller than m.

Response-based. Response-based knowledge distillation uses logits to align the teacher and student. Logits L is the vector obtained after passing through the classification layer. The logits can introduce very high-level information used by teacher to supervise the student.

Again, we introduce a margin m for relaxation to achieve,

$$\mathcal{L}_{res} = \left\lfloor (L_T - L_S)^2 - m \right\rfloor \tag{7}$$

Quality-Based Adaptive Weight (QAW)

In our practice of cross-modal distillation, we find the quality of data can greatly affect the performance. It is not hard to understand this. For example, if the quality of the face image is poor (e.g blurry faces) and the audio data is very high, the teacher (face model) potentially misleads the student. It motivates us to weight the samples by data quality.

There exist many specialized models which can quantify data quality, however, these models are usually computationally expensive. Kim et al. (Kim, Jain, and Liu 2022) prove that there is a high correlation between feature norm and input sample quality. Therefore, we use this simple and effective way to quantify the data quality. In this work, we quantify data quality by using $l2$ norm of the features $\|z_i\|$, and normalizing it to remove the effect of numerical size,

$$Q_i = \frac{\|z_i\| - \mu_z}{\sigma_z} \tag{8}$$

where μ_z and σ_z are the mean and standard deviation of all $\|z_i\|$ in a batch.

The final sample weight is determined by the difference of data quality over two modalities,

$$\Delta Q = Q_T - Q_S \tag{9}$$

$$w_i = \frac{e^{\Delta Q_i}}{\sum_{j=0}^{N} e^{\Delta Q_j}} \tag{10}$$

where N is the batch size.

Experiments

Implementation Details

Input. During training, we use randomly cropped 2 seconds speech segments, without any other data enhancements. 40-dimensional filter-banks (Fbank) with a window of width 25ms and step 10ms are used as the input. For the visual modality, we take 1 frame from each video, crop out the face part, then align it, and finally scale it to 112×112 size as the input of the network.

Datasets. Our model is trained on the VoxCeleb2 (Chung, Nagrani, and Zisserman 2018) dataset and we do the evaluation on the VoxCeleb1 (Nagrani, Chung, and Zisserman 2017) dataset. Both datasets are collected from Youtube. VoxCeleb1 dataset contains 1,251 speakers and over 100,000 utterances, only has data for audio modality. Its original test set contains 37,720 randomly selected pairs, and the hard-set test set contains 552,536 pairs of the same race and gender. VoxCeleb2 dataset contains over a million utterances from over 6,000 speakers and provides both audio and visual modalities. There are no common speakers in the two datasets.

To test the generalizability across datasets, tests were also performed on the CN-Celeb (Fan et al. 2020) dataset. The CN-Celeb dataset contains 11 genres of interviews, singing, movies, etc., in Chinese language. It is very different from the training dataset VoxCeleb, which can effectively test the generalization. Its test set contains 18,849 utterances from 200 speakers and provides 3,484,292 test pairs, which can largely eliminate chance.

Model. For the speaker recognition model, we use X-Vector (Snyder et al. 2018), VGGM (Nagrani, Chung, and Zisserman 2017), ECAPA-TDNN (Desplanques, Thienpondt, and Demuynck 2020) and ResNet34 (with ASP (Okabe, Koshinaka, and Shinoda 2018) for aggregate temporal frames). These models are representatives of the most commonly used and advanced models in speaker recognition. The loss function is a combination of angular prototypical loss (Chung et al. 2020) and cross-entropy loss. The teacher model is a pretrained face recognition model IR-50 taken from (Wang et al. 2021).

Optimization. We use the Adam optimizer with an initial learning rate of 1e-3 decreasing by 25% every 3 epochs and a weight decay of 5e-5. Each batch has 100 speakers, and each speaker has 2 audio utterances. The network is trained for 36 epochs, on an Nvidia RTX 3090 GPU, it takes about 9 hours to train X-Vector and about 2 days to train ResNet34.

Evaluation. Ten 4-second temporal crops are sampled from each test segment for evaluation, and we calculate the distance between all possible pairs (10 × 10 = 100), and use the mean distance as the score. This is the same as (Chung, Nagrani, and Zisserman 2018; Chung et al. 2020). We report two most commonly used evaluation metrics in speaker recognition: the Equal Error Rate (EER) and the minimum value of C_{det}. EER is the rate at which both acceptance and rejection errors are equal. And C_{det} can be calculated by

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}) \tag{11}$$

where we assume a prior target probability P_{tar} of 0.01 and equal weights of 1.0 between misses C_{miss} and false alarms C_{fa}.

To evaluate the robustness under noise, we use the musan dataset (Snyder, Chen, and Povey 2015) to augment our test set. For each piece of test audio, we mix a piece of noise. We set different noise levels based on the decibel gap between

model	EER (O)	minDCF (O)	EER (H)	EER ($\Delta_{db} = 15$)	EER ($\Delta_{db} = 10$)	EER ($\Delta_{db} = 5$)
ResNet34 (w/o dist.)	1.71%	0.205	3.32%	2.78%	3.71%	5.61%
KD (Hinton et al. 2015)	1.70%	0.208	3.60%	2.74%	3.73%	5.52%
PKD (Passalis and Tefas 2018)	1.67%	0.204	3.21%	2.71%	3.57%	5.36%
SP (Tung and Mori 2019)	2.18%	0.259	4.11%	3.50%	4.62%	6.97%
ICKD (Liu et al. 2021)	1.75%	0.209	3.35%	2.88%	3.70%	5.48%
MKD (Feature-based)	**1.51%**	0.178	3.21%	**2.37%**	3.26%	4.97%
MKD (Relation-based)	1.59%	0.187	3.26%	2.63%	3.54%	5.51%
MKD (Response-based)	1.57%	0.193	**3.07%**	2.61%	3.62%	5.53%
MKD (Feature-based) + QAW	1.54%	**0.165**	3.13%	**2.37%**	**3.20%**	**4.77%**

Table 1: Knowledge distillation comparisons using ResNet34. O and H represent the original and hard test set of VoxCeleb1, respectively. The right 3 columns are the test results after adding noises with different levels.

the original speech and the noisy speech. The larger the Δ_{db}, the smaller the noise.

$$\Delta_{db} = db_{audio} - db_{noise} \tag{12}$$

Baseline Results

We first compare the performance of popular speaker recognition methods: X-Vector (Snyder et al. 2018), VGGM (Nagrani, Chung, and Zisserman 2017), ECAPA (Desplanques, Thienpondt, and Demuynck 2020), and ResNet34 (Chung et al. 2020), the input is unified as a 40-dimensional fbank without data augmentation, and the results are shown in Table 2. ResNet34 model achieves the best performance.

method	EER (O)	minDCF (O)	EER (H)
X-Vector (ASP)	7.20%	0.704	13.44%
VGGM (TAP)	4.37%	0.510	7.68%
ECAPA (ASP)	1.75%	0.225	3.67%
ResNet34 (ASP)	**1.71%**	**0.205**	**3.32%**

Table 2: Performance of Baselines. O and H represent the original and hard test sets of VoxCeleb1, respectively.

We then compare the student (speech model) with the teacher (face model). The ResNet34 model is used for student model which achieve best performance in speech. Since the VoxCeleb1 does not have visual modality data which is needed by the teacher, this test was performed on the test set of the VoxCeleb2 dataset. The results in Table 3 show that the pre-trained face recognition model without fine-tuning greatly outperforms the audio model, justifying our assumption that the teacher (face model) is much stronger than the student (speech model).

model	EER	minDCF
Student Model (Audio)	2.89%	0.263
Teacher Model (Visual)	1.97%	0.120

Table 3: Comparisons between the teacher and student on VoxCeleb2 testset.

Comparisons with State-of-the-Art

Note that not all distillation methods can be used for our task. For example, since the spatial locations of visual and speech modalities are not correlated, methods using spatial locations (Zagoruyko and Komodakis 2017) for distillation cannot be applied. Apart from these methods, we compare with some very popular distillation methods (Hinton et al. 2015; Passalis and Tefas 2018; Tung and Mori 2019; Liu et al. 2021) in Table 1. Results show that these methods cannot effectively improve the performance due to the overfitting. Clearly our MKD family can greatly improve the performance by introducing a *loose* distillation which can effectively reduce overfitting.

Apart from ResNet-34, we also test our method using a light-weight X-Vector (Snyder et al. 2018) model in Table 4. Since ICKD (Liu et al. 2021) method cannot work with X-Vector, we only compare with KD (Hinton et al. 2015), PKD (Passalis and Tefas 2018) and SP (Tung and Mori 2019) methods. Results show our method can achieve the greatest performance gains.

model	EER (O)	minDCF (O)	EER (H)
X-Vector	7.20%	0.704	13.44%
KD	7.08%	0.700	12.55%
PKD	6.36%	0.606	10.28%
SP	7.73%	0.751	15.47%
VGSR (Ours)	**6.29%**	**0.585**	**9.94%**

Table 4: Knowledge distillation comparisons using X-Vector. O and H represent the original and hard test set of VoxCeleb1, respectively.

Comparisons with Cross-Dataset Settings

To evaluate the cross-dataset generalizability, the ResNet34 model is trained on VoxCeleb2 and tested on CN-Celeb. These two datasets are very different in terms of video scenes and languages. Since the CN-Celeb dataset is collected from a variety of actual scenarios such as interviews, singing, vlog, etc., it is more difficult and has a higher error rate compared to VoxCeleb as shown in Table 5. Compared to the baseline, our proposed approach VGSR effectively improves the cross-dataset generalizability.

model	EER	minDCF
ResNet34 (w/o dist.)	14.49%	0.759
VGSR (Ours)	**13.03%**	**0.713**

Table 5: Cross-dataset generalizability test. Trained on Voxceleb2, tested on CN-Celeb.

Ablation Study

In this section, we conduct abation study to verify the effectiveness of different components of our methods. We perform this study using ResNet34 on VoxCeleb1.

Effect of projection head hyperparameter α. The face embedding from a pretrained face model might have a big discrepancy against speech embedding. We propose a learnable projection head to reduce this discrepancy with the expectation that the feature from projection head can more easily align with the speech feature. This feature itself can degrade the face recognition performance of the original feature. Thus, we introduce a hyperparameter α in Eq. (2) to balance these two features. From Table 6, if α is too small, it means we mainly rely on the projected feature, the face recognition performance might be degraded greatly; If α is too large, we mainly use the original face embedding, which is very far from speech feature. It will cause difficulty for teacher-student distillation. Neither is the best choice, choosing an intermediate value of 0.6 gives the best results.

α	EER	minDCF
0.2	1.68%	0.210
0.4	1.67%	0.195
0.5	1.52%	0.183
0.6	**1.51%**	**0.178**
0.8	1.68%	0.199

Table 6: Effect of mixing ratio of projection head.

Effect of distillation hyperparameter m. The margin m controls the degree of similarity that the teacher and student should achieve. We use feature-based distillation for this experiment. From Table 7, if m is too large, the constraints are too loose and the supervision is weak; If the m is too small, it may lead to over-fitting. The best performance is achieved by $cos(30°)$.

m	EER	minDCF
0	1.66%	0.188
$cos(10°)$	1.62%	0.180
$cos(20°)$	1.71%	0.197
$cos(30°)$	**1.51%**	**0.178**
$cos(40°)$	1.64%	0.197
$cos(50°)$	1.60%	0.189

Table 7: Effect of margin values.

Effect of MKD and QAW. We conduct evaluations on the original test set of VoxCeleb1 and its noisy version by adding various noises from musan dataset. As shown in Table 1, different types of knowledge can effectively be improved, and the feature-based knowledge can achieve the biggest improvements. In addition, the performance achieved by using QAW is further improved, especially in the case of noises.

Visualization of Features

To understand the features we learned, we visualize the features extracted by the face model, its projection head, and the speech model after applying our MKD and QAW. To achieve these features, we use 200 samples randomly selected from the test set. We conduct dimensionality reduction by t-SNE on these features. From Fig. 3, we can see that there is a large domain gap between the two modalities of face (red) and speech (blue), which is difficult to align them directly. After the mapping of the projection head (green), the distance between face and speech features is reduced, making the teacher-student distillation easier. In addition, the features extracted by the projection head and speech model do not completely overlap because the MKD strategy uses a margin to avoid overfitting. Moreover, based on the annotation of gender (dark), it can be seen that the features mapped by the projection head well indicate gender, which is a shared attribute in face and speech, justifying our assumption that face and speech are highly correlated in feature space.

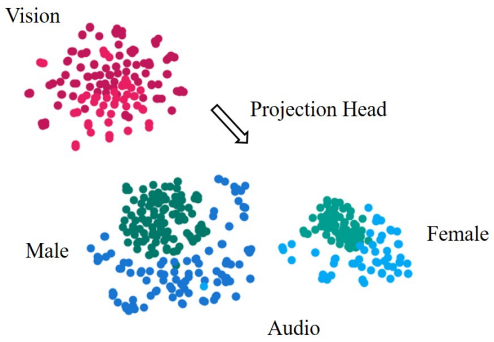

Figure 3: Features extracted by different models. Color red, green and blue represent the features from face model, projection head and speech model respectively. Dark and light colors represent male and female respectively.

Conclusion

In this paper, we propose the VGSR method that utilizes a more discriminative face recognition model as a teacher to guide the student (speech) to improve the performance of speaker recognition. To achieve a promising cross-modal distillation performance between vision-speech modalities, we propose the MKD distillation method and a quality-aware weighting strategy, QAW. Experiments show that our method can effectively transfer discriminative knowledge from face to speech. We hope our cross-modality knowledge transfer strategies can introduce insights into other multi-modal learning tasks.

Acknowledgments

This work was supported by National Natural Science Fund of China (62076184, 61976158, 61976160, 62076182, 62276190), in part by Shanghai Innovation Action Project of Science and Technology (20511100700) and Shanghai Natural Science Foundation (22ZR1466700), in part by Fundamental Research Funds for the Central Universities and State Key Laboratory of Integrated Services Networks (Xidian University).

References

Cai, D.; Wang, W.; and Li, M. 2022. Incorporating Visual Information in Audio Based Self-Supervised Speaker Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 1422–1435.

Chung, J. S.; Huh, J.; Mun, S.; Lee, M.; Heo, H. S.; Choe, S.; Ham, C.; Jung, S.; Lee, B.-J.; and Han, I. 2020. In defence of metric learning for speaker recognition. In *Interspeech*.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech*.

Dehak, N.; Kenny, P. J.; Dehak, R.; Dumouchel, P.; and Ouellet, P. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4): 788–798.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, 3830–3834.

Duarte, A.; Roldan, F.; Tubau, M.; Escur, J.; Pascual, S.; Salvador, A.; Mohedano, E.; McGuinness, K.; Torres, J.; and Giro-i Nieto, X. 2019. Wav2Pix: Speech-conditioned Face Generation Using Generative Adversarial Networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8633–8637.

Fan, Y.; Kang, J.; Li, L.; Li, K.; Chen, H.; Cheng, S.; Zhang, P.; Zhou, Z.; Cai, Y.; and Wang, D. 2020. CN-Celeb: A Challenging Chinese Speaker Recognition Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7604–7608.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision (IJCV)*, 129(6): 1789–1819.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2827–2836.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Inoue, N. 2021. Teacher-assisted mini-batch sampling for blind distillation using metric learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4160–4164. IEEE.

Kamachi, M.; Hill, H.; Lander, K.; and Vatikiotis-Bateson, E. 2003. 'Putting the Face to the Voice': Matching Identity across Modality. *Current Biology*, 13(19): 1709–1714.

Kim, M.; Jain, A. K.; and Liu, X. 2022. AdaFace: Quality Adaptive Margin for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, L.; Huang, Q.; Lin, S.; Xie, H.; Wang, B.; Chang, X.; and Liang, X. 2021. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8271–8280.

Nagrani, A.; Albanie, S.; and Zisserman, A. 2018. Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: a large-scale speaker identification dataset. In *Interspeech*.

Oh, T.-H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; and Matusik, W. 2019. Speech2Face: Learning the Face Behind a Voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Okabe, K.; Koshinaka, T.; and Shinoda, K. 2018. Attentive statistics pooling for deep speaker embedding. In *Interspeech*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3967–3976.

Passalis, N.; and Tefas, A. 2018. Probabilistic knowledge transfer for deep representation learning. *CoRR, abs/1803.10837*, 1(2): 5.

Peng, B.; Jin, X.; Liu, J.; Li, D.; Wu, Y.; Liu, Y.; Zhou, S.; and Zhang, Z. 2019. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5007–5016.

Qian, Y.; Chen, Z.; and Wang, S. 2021. Audio-visual deep neural network for robust person verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1079–1092.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representation (ICLR)*.

Sarı, L.; Singh, K.; Zhou, J.; Torresani, L.; Singhal, N.; and Saraf, Y. 2021. A multi-view approach to audio-visual speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6194–6198. IEEE.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In

Proceedings of the IEEE conference on computer vision and pattern recognition, 815–823.

Smith, H. M. J.; Dunn, A. K.; Baguley, T.; and Stacey, P. C. 2016. Concordant Cues in Faces and Voices: Testing the Backup Signal Hypothesis. *Evolutionary Psychology*, 14(1): 1474704916630317.

Snyder, D.; Chen, G.; and Povey, D. 2015. MUSAN: A Music, Speech, and Noise Corpus. ArXiv:1510.08484v1, .

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5329–5333. IEEE.

Tao, R.; Das, R. K.; and Li, H. 2020. Audio-visual Speaker Recognition with a Cross-modal Discriminative Network. In *Interspeech*.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1365–1374.

Wang, Q.; Zhang, P.; Xiong, H.; and Zhao, J. 2021. Face. evoLVe: A High-Performance Face Recognition Library. *arXiv preprint arXiv:2107.08621*.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representation (ICLR)*.

Zhang, L.; Chen, Z.; and Qian, Y. 2021. Knowledge Distillation from Multi-Modality to Single-Modality for Person Verification. *Proc. Interspeech 2021*, 1897–1901.