INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Knowledge Distillation from Multi-Modality to Single-Modality for Person Verification

Leying Zhang, Zhengyang Chen, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence, AI Institute X-LANCE Lab,
Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{zhangleying, zhengyang.chen, yanminqian}@sjtu.edu.cn

Abstract

Voice and face are two important biometric characteristics that can be used for person identity verification. Previous works have proved the strong complementarity between audio and visual modalities in person verification tasks that multi-modality system can achieve significant performance improvement compared to single-modality system. However, due to the limitations in the real world, it is hard to access both audio and visual data at the same time. In this paper, we investigate several strategies to distill the knowledge from a multi-modality system and transfer it to the single-modality system in a teacher-student mode. We applied the knowledge distillation at three different levels: label level, embedding level, and distribution level. All the experiments are based on the VoxCeleb dataset. The results show that the visual single-modality system achieves 10% EER (equal error rate) improvement on the VoxCeleb1 evaluation set using our proposed knowledge distillation method. Besides, the improvement on the audio system is only reflected on part of the evaluation trials, and we give a detailed analysis for this phenomenon.

Index Terms: person verification, knowledge distillation, multi-modality, audio and visual

1. Introduction

Recent years have witnessed the wide application of multi-modality systems, especially audio-visual systems. Face and voice serving as two essential aspects of human expression have gained extensive attention of researchers. Previous studies in neural science have demonstrated that visual and auditory neural signals interact in the cognitive process [1, 2] and researchers in computer science also studied cross-modal biometric matching and correlation between audio and visual information [3, 4, 5]. It has been observed that image processing capabilities in lip reading assist speech recognition systems [6] and additional facial information helps speech segmentation at the cocktail party [7].

There is no doubt that the multi-modality system has great superiority over the single-modality system because of the complementarity between different modalities. For person verification task, researchers have shown that the multi-modal fusion at the decision [8, 9, 10] and embedding level [11, 12, 13] can both significantly improve the system's performance. However, in some specific scenes of real life, it is always hard to access one's facial and voice information simultaneously. For example, the face recognition system may fail in the case of a person wearing a mask and the speaker verification system can't verify one identity in the case of strong background noise. To further leverage the multi-modality system's superiority, it could be better if we can transfer the knowledge from the multi-modality system to the single-modality system.

Knowledge distillation (KD) [14] has been proposed to transfer information learned from one model to another, and it is often characterized as a ”Teacher-Student” (TS) learning mode. Researchers have widely used knowledge distillation methods in the model compression task [15, 16], where a large model with better performance acts as a teacher to teach a small model and help the small model achieve further improvement. In this paper, we consider the multi-modality system as the teacher to transfer its knowledge to the single-modality system.

To implement knowledge distillation based on a multi-modality system, we first trained an audio-visual system following our previous work [12, 13]. Then the audio-visual system teaches the single-modality system at three different levels:

- Embedding level knowledge distillation: Directly use the embedding from the audio-visual system to guide the single-modality system optimization based on cosine similarity.
- Label level knowledge distillation: Here, the person class posterior predicted by the audio-visual system are used as the auxiliary label for the single-modality system training. The Kullback-Leibler divergence is used to calculate the loss.
- Distribution level knowledge distillation: Here, instead of forcing the embedding and output posterior mentioned above exactly matching for each TS pair, we use the Maximum Mean Discrepancy (MMD) [17] to constrain the embedding distribution of single-modality system similar to the multi-modality one's.

The rest of the paper is organized as follows. Section 2 introduces the related work about teacher-student knowledge distillation, cross modality distillation, and multi-modality fusion. Section 3 will give detailed description about our multi-modality knowledge distillation methods from three perspectives. Experimental setups and result analysis will be shown in section 4 and 5. Finally, we conclude in section 6.

2. Related Work

Teacher-Student Knowledge Distillation: Knowledge distillation is originally proposed in [18] and popularized in [14]. It has received rapidly increasing attention from the academic and industrial community [19]. Using this method, relatively small networks can achieve good recognition results under the guidance of more complicated models [16, 15]. The standard knowl-

[†] corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-2119

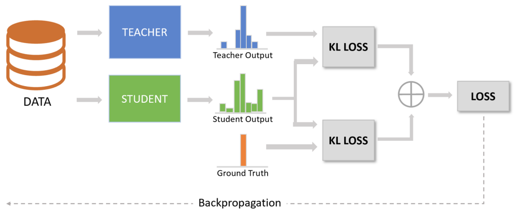

Figure 1: *An illustration of standard knowledge distillation*

edge distillation process is shown in Fig. 1, where the student system is trained using ground truth labels and pseudo labels of teacher system simultaneously. Researchers further studied correlations within instances [20] and proposed probabilistic distributions matching method [21] to improve knowledge distillation efficiency. In this paper, we explore the knowledge distillation from more perspectives compared with the standard one.

Cross Modality Distillation: Cross modality knowledge distillation is a hot topic, where information from one modality is used to help train another modality. In [22], a Conditional Generative Adversarial Network (CGAN) was proposed to distill knowledge from sensor data and enhance low-resolution target detection. Besides, as described in [23], depth side information contributes in RGB object detection. For speaker diarization tasks, researchers in [24] proposed two methods, namely target embedding transfer and clustering structure transfer, for improving system performance by utilizing face embeddings. However, most researches are limited to cross-modal knowledge transferring, the study of using excellent performance of multi-modality system to teach single-modality systems is paid less attention.

Multi-Modality Fusion: Previous work has proved the efficiency of audio-visual fusion systems in person verification task. There are two popular methods to fuse the information from audio and visual modality in person verification task, the decision-level fusion [8, 9, 10] and embedding-level fusion [11, 12]. Both fusion systems can boost the verification system's performance. To better distill the knowledge from the audio-visual fusion system, we use the best embedding-level fusion system in [12] as the teacher system.

3. Multi-Modality Knowledge Distillation

3.1. Single- and Multi-Modality System

As shown in Figure 2, we denote audio and visual inputs as $\mathbf{x}_a$ and $\mathbf{x}_v$ respectively including N trial pairs of C speakers. Then, the audio embedding extractor F_a and visual embedding extractor F_v can map the input $\mathbf{x}_a$ and $\mathbf{x}_v$ to person embedding $\mathbf{e}_a$ and $\mathbf{e}_v$ separately, where $\mathbf{e}_a = F_a(\mathbf{x}_a)$ and $\mathbf{e}_v = F_v(\mathbf{x}_v)$. To leverage the multi-modality information, another audio-visual fusion system, F_{av}, is used to fuse the audio and visual embedding to the fusion embedding $\tilde{\mathbf{e}}_{av}$.

The single- and multi-modality systems are all optimized using additive angular margin (AAM) loss [25]. To calculate the AAM loss, a projection matrix $\mathbf{W} \in \mathbb{R}^{C \times d}$ is introduced, where C is person classification number and d is embedding dimension. Each column of $\mathbf{W}$ can be used to represent the embedding center of one person. Using θ_j^i to denote the angle between embedding $\mathbf{x}_i$ and j^{th} column of $\mathbf{W}$, the AAM loss can be formulated as:

$$\mathcal{L}^{AAM} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(\cos(\theta^i_{y^i}+m))}}{e^{s(\cos(\theta^i_{y^i}+m))} + \sum_{j=1, j\neq y^i}^{C} e^{s(cos(\theta^i_j))}} \tag{1}$$

where m and s are the margin and scale hyper-parameter in AAM loss, and y_i is the ground truth label for i^{th} sample.

3.2. Label-level Knowledge Distillation

Based on the projection matrix of AAM introduced in the last section, the person class posterior for each input embedding $\mathbf{e}$ can be calculated as Eq. 2, where σ is softmax function and T is temperature fixed in our experiment.

$$\mathbf{P} = \sigma(\frac{\|\mathbf{W}\|\|\mathbf{e}\|}{T}) \tag{2}$$

Using the person posterior from single- and multi-modality system, we do the label-level knowledge distillation based on Kullback-Leibler divergence and loss can be formulated as:

$$\mathcal{L}^{KL} = -\sum_{i=1}^{N}\sum_{j=1}^{C} \tilde{\mathbf{P}_j^i} \log \mathbf{P}_j^i \tag{3}$$

3.3. Embedding-level Knowledge Distillation

In this section, we explore the knowledge distillation at the embedding level. The AAM loss introduced in section 3.1 optimize the person embedding in a hyper-sphere space. Thus, cosine distance can be a reasonable similarity metric between the embeddings from teacher and student system. The knowledge distillation loss based on cosine distance is written as:

$$\mathcal{L}^{COS} = 1 - \sum_{i=1}^{N} \frac{\tilde{\mathbf{e}^i} \cdot \mathbf{e}^i}{\|\tilde{\mathbf{e}^i}\|\|\mathbf{e}^i\|} \tag{4}$$

where $\tilde{\mathbf{e}^i}$ represents the embedding extracted from the teacher network for the i^{th} sample, $\mathbf{e}^i$ means the embedding computed by the student network.

3.4. Distribution-level Knowledge Distillation

Both the label- and embedding-level knowledge distillation impose a strong constraint between the embeddings or posteriors from teacher and student models. Intuitively, although there is some association between one person's facial and voice information, the gap between different modalities does exist. Here, we introduce a weaker constraint when doing the teacher-student knowledge distillation that we guide the student model learn the embedding distribution from the teacher. And we use the Maximum Mean Discrepancy (MMD) to achieve this goal.

MMD is a distance on the space of probability measures described in [17], which is widely used in transfer learning but also efficient in knowledge distillation [20, 26, 27]. It helps to analyze and compare distributions so as to determine if two samples are drawn from different distributions by projecting the sample distribution on Reproducing Kernel Hilbert Spaces. Let $X, Y \in \mathbb{R}^{m\times n}$ be two observations i.i.d., and let $k(x, y)$ be the kernel function. By applying unbiased empirical estimate, we get squared unbiased MMD in Eq. 5.

$$\begin{aligned}\mathcal{L}^{MMD}[\mathcal{F}, X, Y] = \|\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) \\ + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)\|^{\frac{1}{2}}\end{aligned} \tag{5}$$

In this paper, we choose Gaussian kernel defined in Eq. 6.

$$k(x, y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}} \tag{6}$$

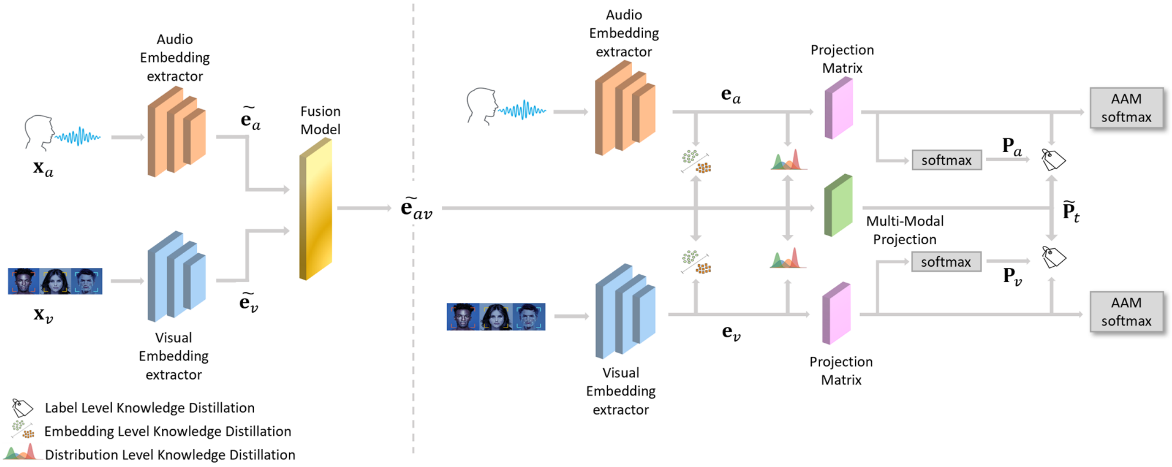

Figure 2: *Multi-modality knowledge distillation framework.* $\mathbf{e}, \mathbf{P}$ *represent embedding and class posterior. Symbols with* $\sim$ *are for teacher system.* ***Left:*** *multi-modality teacher system and its embedding extraction process.* ***Right:*** *teacher-student method at three different levels for the single-modality student system.*

We utilize maximum mean discrepancy with 5 Gaussian kernels as loss function $\mathcal{L}^{MMD}[\mathcal{M}, \mathcal{N}]$. $\mathcal{M}$ is a set of teacher's embedding from one batch, and $\mathcal{N}$ is a set of student's embedding from the corresponding batch. In the MMD calculation, all the embeddings are L2-normalized to ensure the consistent embedding scale between teacher and student.

With the above described three-level knowledge-distillation approaches, the final optimization objective for knowledge distillation from multi-modality to single-modality is shown as Eq. 7. The distillation losses at different levels are weighted summed with the main AAM optimization objective. We will explore their separate and combined effects by altering the hyper-parameters α, β, and γ.

$$\mathcal{L}^{KD} = \mathcal{L}^{AAM} + \alpha\mathcal{L}^{KL} + \beta\mathcal{L}^{COS} + \gamma\mathcal{L}^{MMD} \tag{7}$$

4. Experimental setups

4.1. Dataset

In our experiments, we use the audio and visual data from VoxCeleb which contains two parts, VoxCeleb1 [28] and VoxCeleb2 [29]. We use the dev part of VoxCeleb2 in the training process, comprising 5994 speakers. The whole VoxCeleb1 dataset is used for evaluation. Three official trials Vox1-O, Vox1-E, and Vox1-H for VoxCeleb1 are used to report the results.

4.2. System Configuration

Single-Modality System: The 40-dimensional Fbank with 25ms window length and 10ms hop length is used as the input audio feature. We randomly sample a segment between 200 to 400 frames from each utterance during the training process. Data augmentation is not used for audio data. For visual data, the facial image extracting process is the same as the pipeline described in [12]. Data augmentation is not used for visual data. We apply ResNet34 described in [30] as the audio single-modality network and ResNet34 described in [31] as the visual single-modality network. The embedding dimension of both networks is 512. AAM loss is used to optimize both single-modality systems, where margin and scale are set to 0.2 and 32, respectively.

Multi-Modality System: The embeddings from the audio and visual systems are pre-extracted for each video segment. We use the Gated Multi-Modal Fusion (GATE) multi-modality system proposed in [12] to fuse the audio and visual embedding. AAM loss is also used to train the fusion system, where margin and scale are set to 0.7 and 32, respectively.

Knowledge Distillation Training Details: For knowledge distillation training, we directly used the embeddings extracted from the audio-visual system without updating the parameters of fusion system. Especially, for label-level knowledge distillation, the projection matrix of AAM from audio-visual system is used to calculate the teacher system's person posterior and is fixed during training. Besides, the temperature T of KL Divergence loss is fixed to 0.03125.

5. Results and Analysis

To transfer knowledge from multi-modality system to the visual and audio system respectively, we investigated and compared different distillation strategies in our experiment. We performed cosine distance scoring for evaluation. Results and analysis will be presented in this section.

5.1. Results of Single and Multi-Modality Systems on the voxceleb1

Table 1: *Results (EER %) of Single and Multi-Modality Systems*

Modality	**Vox1-O**	**Vox1-E**	**Vox1-H**
Audio	1.792	1.704	2.964
Visual	1.299	0.987	1.483
Audio-Visual	**0.514**	**0.375**	**0.555**

The results of single and multi-modality systems are shown in Table 1. In our experiment, we further optimize the visual system's training strategy compared to the ones in our previous work [12]. The visual system performs even better and we get a stronger fusion system. Besides, the fusion system significantly outperforms both single-modality systems, showing the complementary ability between audio and visual modalities.

5.2. Knowledge Distillation to Visual System

In this section, we will explore and compare different knowledge distillation methods for the visual system. Table 2 shows corresponding results. Compared to the baseline system, all the visual systems after knowledge distillation from the audio-

Table 2: *Results (EER %) comparison of visual system using different losses. The hyper-parameters in Eq. 7 are tuned a bit, and $\mathcal{L}^{AAM}$ is used in all the following results.*

Loss	Vox1-O	Vox1-E	Vox1-H
Visual-Baseline	1.299	0.987	1.483
$\mathcal{L}^{COS}$	1.294	0.968	1.427
$\mathcal{L}^{KL}$	1.257	0.905	**1.331**
$\mathcal{L}^{MMD}$	**1.140**	**0.903**	1.341
$\mathcal{L}^{COS}$+$\mathcal{L}^{KL}$	1.288	0.938	1.386
$\mathcal{L}^{COS}$+$\mathcal{L}^{MMD}$	1.278	0.928	1.355
$\mathcal{L}^{KL}$+$\mathcal{L}^{MMD}$	1.235	0.903	1.346

visual system achieve further improvement. Inconsistent with the results in other KD based model compression work [15], the cosine similarity with strong constraint performs worse than the KL and MMD constraint. This phenomenon is reasonable because there is a large gap between audio and visual information that we can't enforce the embeddings from teacher and student exactly matching. Besides, the MMD loss achieves the best performance by simply matching the teacher and student embedding distribution in a batch. Finally, we explored combining different knowledge distillation methods together, and the results didn't show another improvement, which further confirms that strong constraint is not suitable for multi-modality knowledge distillation.

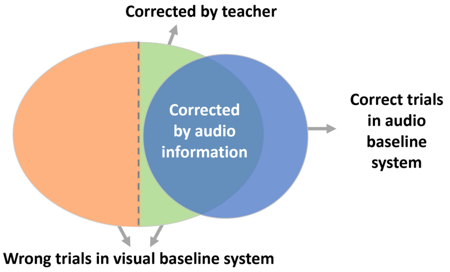
Corrected by teacher
Corrected by audio information
Correct trials in audio baseline system
Wrong trials in visual baseline system

Figure 3: *The trial pair results distribution Venn diagram for visual system based on Vox1-E.*

To further prove the effects of multi-modality knowledge distillation, we analyze the knowledge learned by the visual system from the teacher based on different system's score on Vox1-E. As the green part shown in Figure 3, there are 5737 trial pairs which are originally misjudged by the single-modality visual baseline system and then corrected by the knowledge distillation. Surprisingly, 4877 (85% of 5737) trial pairs of them were correctly predicted by the audio system baseline, which indicates that the teacher system leverages the audio information to help the visual system.

5.3. Knowledge Distillation to Audio System

Here, the multi-modality knowledge distillation is implemented on the audio system. Table 3 summarizes the corresponding results. Audio-systems with knowledge distillation all achieves great improvement compared to baseline on Vox1-O and the improvements on Vox1-E and Vox1-H is not so obvious. Embedding-level distillation based on cosine similarity performs worst among all the knowledge distillation methods, which is consistent with our findings in the visual system.

Besides, although both student models are under the guidance of the same teacher, according to the above results, the audio model learns worse than the visual model. We also analyze knowledge learned by the audio student system. There are 1945 trial pairs on Vox1-E initially misjudged by audio baseline system and then corrected by the knowledge distillation. Consistent with the situation shown in Fig 3, the visual system baseline correctly predicted 1862 trial pairs of them. However, the total amount of knowledge learned by audio student system is much smaller than that learned by visual student system, and we explore some clues to explain this phenomenon. Previous researches [32, 33] have shown that a model with well performance is not necessarily a good teacher if the performance gap between teacher and student model is too large.

Table 3: *Results (EER %) comparison of audio system using different losses. The hyper-parameters in Eq. 7 are tuned a bit, and $\mathcal{L}^{AAM}$ is used in all the following results.*

Loss	Vox1-O	Vox1-E	Vox1-H
Audio-Baseline	1.792	1.704	2.964
$\mathcal{L}^{COS}$	1.707	1.690	2.960
$\mathcal{L}^{KL}$	**1.671**	1.686	**2.935**
$\mathcal{L}^{MMD}$	1.718	**1.680**	2.947
$\mathcal{L}^{COS}$+$\mathcal{L}^{KL}$	1.691	1.692	2.963
$\mathcal{L}^{COS}$+$\mathcal{L}^{MMD}$	1.718	1.692	2.943
$\mathcal{L}^{KL}$+$\mathcal{L}^{MMD}$	1.697	1.684	2.950

To further explore the relationship between our single-modality system and multi-modality system, we used the Pearson Correlation Coefficient (PCC) [34]. A larger PCC means stronger correlation, a value of 0 implies no linear correlation between the variables. The PCC between visual embeddings before fusion and audio-visual embeddings equals 0.079, and the PCC between audio embeddings before fusion and audio-visual embeddings equals 0.007. Moreover, when utilizing cosine similarity for knowledge distillation, after teacher-student training, the cosine distance loss of visual system has dropped to 0.4217, but the cosine distance loss of audio system is 0.583, which means it is more difficult for audio system to approximate embedding of teacher system. Phenomenons shown above indicate that the multi-modal system in our experiment may be not a good teacher for audio system and we will find a better way to help the multi-modality teacher to teach audio system.

6. Conclusions

In this paper, we developed several knowledge distillation methods to distill knowledge from the multi-modality system to the single-modality system, including label-level, embedding-level and distribution level knowledge distillation approaches. Results show that the visual system benefits a lot from the distilled knowledge and achieves 10% relative improvement on the VoxCeleb1 evaluation set. For audio student system, the improvement is mainly reflected in some evaluation trials and we analyzed the reasons in detail. In the future, we will explore more effective teacher-student method to distill knowledge from audio-visual system to the audio system.

7. Acknowledgements

This work was supported by the China NSFC projects (No. 62071288 and No. U1736202). Experiments have been carried out on the PI supercomputers at Shanghai Jiao Tong University.

8. References

[1] K. v. Kriegstein, A. Kleinschmidt, P. Sterzer, and A.-L. Giraud, "Interaction of face and voice areas during speaker recognition," *Journal of cognitive neuroscience*, vol. 17, no. 3, pp. 367–376, 2005.

[2] L. W. Mavica and E. Barenholtz, "Matching voice and face identity from static images." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 2, p. 307, 2013.

[3] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436.

[4] S. Horiguchi, N. Kanda, and K. Nagamatsu, "Face-voice matching using cross-modal embeddings," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1011–1019.

[5] R. Tao, R. K. Das, and H. Li, "Audio-visual speaker recognition with a cross-modal discriminative network," *arXiv preprint arXiv:2008.03894*, 2020.

[6] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[7] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[8] J. Luque, R. Morros, A. Garde, J. Anguita, M. Farrus, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, and J. Hernando, "Audio, video and multimodal person identification in a smart room," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 258–269.

[9] E. Erzin, Y. Yemez, A. M. Tekalp, A. Erçil, H. Erdogan, and H. Abut, "Multimodal person recognition for human-vehicle interaction," *IEEE MultiMedia*, vol. 13, no. 2, pp. 18–31, 2006.

[10] J. S. Chung, B.-J. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," *arXiv preprint arXiv:1906.10042*, 2019.

[11] S. Shon, T.-H. Oh, and J. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3995–3999.

[12] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[13] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on voxceleb," *Proc. Interspeech 2020*, pp. 2252–2256, 2020.

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[15] S. Wang, Y. Yang, T. Wang, Y. Qian, and K. Yu, "Knowledge distillation for small foot-print deep speaker embedding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6021–6025.

[16] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for bert model compression," *arXiv preprint arXiv:1908.09355*, 2019.

[17] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[18] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 535–541.

[19] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *arXiv preprint arXiv:2006.05525*, 2020.

[20] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.

[21] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[22] S. Roheda, B. S. Riggan, H. Krim, and L. Dai, "Cross-modality distillation: A case for conditional generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2926–2930.

[23] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 826–834.

[24] N. Le and J.-M. Odobez, "Improving speaker turn embedding by crossmodal transfer learning from face embedding," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 428–437.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[26] M. Wang, R. Liu, N. Hajime, A. Narishige, H. Uchida, and T. Matsunami, "Improved knowledge distillation for training fast low resolution face recognition model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[27] P. Zhou, L. Mai, J. Zhang, N. Xu, Z. Wu, and L. S. Davis, "M2kd: Multi-model and multi-level knowledge distillation for incremental learning," *arXiv preprint arXiv:1904.01769*, 2019.

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[29] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[30] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4794–4802.

[33] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[34] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.