



VoxBlink2: A 100K+ Speaker Recognition Corpus and the Open-Set Speaker-Identification Benchmark

Yuke Lin^{1,3}, Ming Cheng^{1,3}, Fulin Zhang², Yingying Gao², Shilei Zhang^{*2}, Ming Li^{*1,3}

¹School of Computer Science, Wuhan University, Wuhan, China

²China Mobile Research Institute, Beijing, China

³Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Duke Kunshan University, Kunshan, China

Abstract

ming.li@whu.edu.cn

In this paper, we provide a large audio-visual speaker recognition dataset, VoxBlink2, which includes approximately 10M utterances with videos from 110K+ speakers in the wild. This dataset represents a significant expansion over the VoxBlink dataset, encompassing a broader diversity of speakers and scenarios by the grace of an optimized data collection pipeline. Afterward, we explore the impact of training strategies, data scale, and model complexity on speaker verification and finally establish a new single-model state-of-the-art EER at 0.170% and minDCF at 0.006% on the VoxCeleb1-O test set. Such remarkable results motivate us to explore speaker recognition from a new challenging perspective. We raise the Open-Set Speaker-Identification task, which is designed to either match a probe utterance with a known gallery speaker or categorize it as an unknown query. Associated with this task, we design concrete benchmark and evaluation protocols. The data and model resources can be found in <http://voxblink2.github.io>.

Index Terms: Speaker Verification, Dataset, Multi-modal.

1. Introduction

Speaker recognition has been widely studied over the past decades, resulting in tremendous performance improvements. Although numerous efforts have been focused on various model structures[1, 2, 3, 4, 5] under complex application scenarios[6, 7, 8, 9], there is still a significant gap towards the requirement of commercial applications. Given the prevailing trend of large models across different domains, we anticipate that large scaled datasets and models performance dramatically.

The variability and quality of data play essential roles in the development of robust speaker recognition systems. The VoxCeleb[10, 11] is currently the most popular database for speaker recognition. Nevertheless, when compared to datasets for facial recognition, the disparity in size is profound, spanning two orders of magnitude. The VTL [12] includes a huge amount of utterances from 100K+ speakers, but only a small version[13] of pure audio with 5,040 speakers is publicly available. The VoxBlink dataset[14] introduces a novel, highly scalable data-mining method for data collection. However, it did not bring about a qualitative improvement in data volume. Hence, we have refined the data collection pipeline of VoxBlink and expanded the scope of retrieval, thereby aggregating a much bigger scale audio-visual dataset for many possible applications. Moreover, we discover that further increasing the size of data and the complexity of models can achieve **state-of-the-art** results, which drives us to explore a more challenging scenario.

* Corresponding authors: Shilei Zhang, Ming Li.

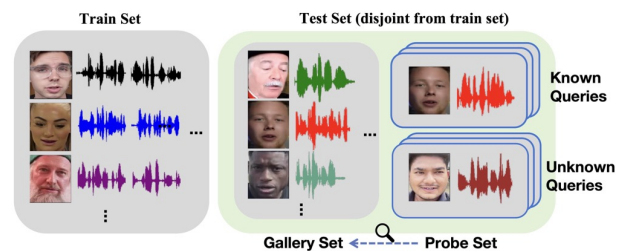


Figure 1: The outline of the Open-Set Speaker-Identification task. The evaluation protocol requires a gallery set for speaker enrollment. Then, the test queries must be linked to the speaker in the gallery as known queries or rejected as unknown queries based on the similarities scores and a pre-defined threshold.

While the verification task requires only a 1: 1 comparison, the identification problem requires 1: N comparisons – of a probe sample with templates from identities in a gallery. Depending on the gallery size N, finding the correct answer can be much more challenging than simply performing a correct 1: 1 comparison. Moreover, the unseen individual in the probe set should not be linked to the gallery identities (like the entrance guard, etc.) in many scenarios. Therefore, we raise the Open-Set Speaker-Identification (OSSSI) benchmark based on the VoxBlink-clean set, incorporating over ten thousand speakers in evaluation for the first time. We adopt the Detection and Identification Rate at the False Alarm Rate (DIR@FAR)[15], which works under an m: n open-set protocol, with the ability to identify enrolled speakers and reject unseen identities. In general, our contribution can be summarized as follows:

- We optimize the data collecting pipeline and release a large audio-visual speaker recognition dataset at the scale of 100K+ speakers.
- We explore different training strategies with different model sizes and data volumes to show the trend, and achieve the state-of-the-art system performance on the Vox1 test set.
- We propose a new Open-Set Speaker-Identification benchmark as well as evaluation protocols and baselines.

2. VoxBlink2 Dataset

2.1. Data Description

The VoxBlink2 corpus is composed of 9,904,382 high-quality utterances and their corresponding video clips, sourced from 111,284 users on YouTube. To our best knowledge, it is the largest publicly available audio-visual dataset for speaker recognition. Unlike the VoxBlink[14], which only utilizes short video segments, we extract the initial minute of long-duration

user-uploaded videos, significantly expanding the diversity and scenarios of data. Other information is detailed in Tab.1.

Table 1: *The statistics for the VoxBlink2 dataset compared with the VoxCeleb2 and the VoxBlink. Utt and Dur mean utterance and duration, respectively.*

Dataset	VoxCeleb2	VoxBlink	VoxBlink2
# of SPKs	5,994	38,065	111,284
# of videos	145,569	372,084	2,097,062
# of hours	2,442	2,135	16,672
# of utterances	1,092,009	1,455,190	9,904,382
Avg # of Utts per spk	185	38.23	89
Avg # of Dur. per Utt	7.8	5.3	6.0
Avg # of Span(days) per spk	-	440	786

2.2. Data mining

2.2.1. Collection Pipeline

While the data mining process for VoxBlink2 follows a pipeline similar to that of VoxBlink, several modifications have been implemented to enhance data quality further. The stem pipeline can be outlined as follows:

Step I. Candidate Collection. Considering the impact of language diversity on speaker recognition systems, we compile a long keyword list spanning 18 languages for user retrieval. Then, we collect over 6 million 1-min videos from Youtube users who utilize their photos as avatars. It is noteworthy that we intentionally avoid duplicate users with the VoxBlink and duplicate recordings with the VoxCeleb1&2.

Step II. Frame Extraction & Face Detection. In pursuit of higher quality and efficiency, we employ a high frame rate (25 fps) for frame extraction and utilize the MobileNet[16] to detect facial movements. The 1-person video tracks are generated by setting a threshold of the minimum Intersection Over Union (IOU) value between two consecutive units, ensuring that each facial track includes only one person.

Step III. Face Recognition. After the face detection, we identify the faces along the video track by our pre-trained ArcFace classifier, which is introduced in 2.2.2. Adopting the identification approach rather than verification enhances data purity.

Step IV. Active Speaker Detection & Overlap Speech Detection. To mitigate the inclusion of silent and overlapped segments, we integrate an audio-visual speaker diarization model[17] and an overlap detection model[18] into our pipeline. These models enable the partitioning of active speech segments and eliminating overlapping speech segments, respectively.

2.2.2. Face Classifier Training

The accuracy of open-set 1:1 verification is often influenced by inter-domain differences, leading to some error labels in datasets constructed using face verification methods such as VoxBlink. As illustrated in Fig. 2, we have introduced a **supplementary branch** (highlighted in blue) dedicated to training a classifier, following the outlined procedures:

Coarse Frame Extraction & Face Detection. In contrast to Step II described in Sec.2.2.1, we utilize a relatively low frame rate to capture frames from all candidate videos. Subsequently, frames featuring single-person appearance are exclusively detected by the mobilenet[16] for face verification.

Face Verification. The ResNet-IRSE50 model [19] is employed to extract face embeddings from the obtained facial images and compute 1:1 similarity scores with the template embedding from the candidate’s avatar photo.

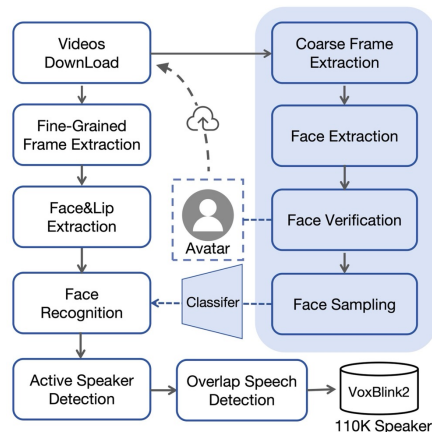


Figure 2: *The outline of the data collection pipeline. The modules highlighted in blue(right) are designed to extract faces from candidates for better accuracy.*

Face Sampling. We use cosine scores to weighted-sample faces from each speaker, with the maximum number of faces capped at 10 per speaker. In short, faces exhibiting higher cosine similarity to the avatar embedding are prioritized for selection for training.

Ultimately, we concatenate ArcFace at the end of the ResNet-IRSE50 encoder and proceed to train the encoder with the classifier by approximately 200K candidate face data collected from face sampling. Furthermore, the VoxBlink2 dataset demonstrates a substantial increase in accuracy, reaching 92%, compared to the 72% accuracy achieved by the VoxBlink dataset, as verified through manual assessment of a randomly sampled subset of 50 speakers.

3. Open-Set Speaker-Identification

Traditional 1:1 verification tasks are typically deployed on terminal devices, whereas the m: n open-set identification offers an opportunity to integrate speaker recognition systems into practical access control systems. Despite this potential, the exploration of this topic in speaker recognition remains limited, partly due to the absence of large-scale evaluation protocols. Although [20] and [21] have made some endeavors, the open-source evaluation sets are not publicly accessible to academia, and the scale of data is relatively limited. Motivated by the methodology proposed in [15, 22] for face recognition, our proposed protocol aims to establish a standardized framework for evaluating speaker recognition systems in open-set identification scenarios, with the ability to recognize enrolled speakers and reject unprivileged speakers.

3.1. Evaluation Protocols

As depicted in Fig.1, the evaluation protocols involve two different sets: gallery set (G) and probe set (P). The gallery set (G) is constructed with an equal number of enrollment utterances for each known identity, essentially serving as a knowledge base. On the other hand, P is dedicated to querying G and is divided into two categories: Known Queries (K) and Unknown Queries (U). Here, we adopt the VoxBlink-clean set (1,028,095 utterances from 18,381 speakers) and fabricate the following protocols in terms of different application scenarios, which is shown in Tab.2. Furthermore, we set for each protocol with 1, 3, and 5 enrollment utterances per speaker in G , therefore adding up to $3*3=9$ different evaluation protocols.

Table 2: Basic information of three evaluation protocols. The numbers denote the numbers of speakers in gallery, known and unknown query sets. S, M, L indicate Small, Medium, Large.

	G	K	U	Authentication Scenarios
VB-Eval-S	60	30	30	Exam room
VB-Eval-M	600	300	300	Office building
VB-Eval-L	6,000	3,000	3,000	Major events

3.2. Evaluation Metrics

The OSSI can be measured by the Detection and Identification Rate (DIR) and False Alarm Rate (FAR)[15]. Moreover, we adopt the DIR@FAR to recognize known speakers while maintaining a fixed FAR threshold.

Initially, upon obtaining a feature extractor trained on the dataset, speaker embeddings are simultaneously extracted from G and P , where $(P = K \cup U)$. When a known probe p is presented to a system, the similarity scores between p and all samples in G are computed and sorted. A probe p has rank n if $s(p, G_p)$ is the n -th largest speaker embedding similarity score, here G_p represents the matched speaker with p in G .

Finally, for a given pre-defined similarity threshold θ and a rank n , the DIR can be derived by Eq.1. In our evaluation framework, the n is set to 1 to calculate the top-match speaker, and we use the cosine-similarity for similarity calculation.

$$DIR(\theta, n) = \frac{|rank(p) \geq n \wedge sim(p, G_p) \geq \theta; p \in K|}{|K|} \quad (1)$$

The False Alarm Rate (FAR) serves as a measure of the system’s ability to discern and reject unknown queries, typically considered impostors. A false alarm event occurs when the top match score for an imposter in U is higher than θ . Assuming that G_p symbolises the most-matched speaker with p in G , the FAR can be computed by:

$$FAR(\theta) = \frac{|sim(p, G_p) \geq \theta; p \in U|}{|U|} \quad (2)$$

The optimal system should have a 1.0 DIR and a 0.0 FAR, indicating perfect detection and identification of all individuals in the probe set without any false alarms. However, in real-world systems, there is a trade-off between the DIR and the FAR. This trade-off is influenced by varying threshold values (θ) that can be adjusted to meet specific FAR requirements, and visualizing this trade-off is often done through a Receiver Operating Characteristic (ROC) curve. For some well-defined authentication scenarios with pre-defined FAR, we can use the DIR at a particular FAR (DIR@FAR) value as a metric.

4. Experimental Settings

Data Usage. Our experiments primarily utilize the VoxBlink2 (VB2) and VoxCeleb2 (VC2) datasets for training, while evaluations are conducted on the VoxCeleb1 test set for the speaker recognition task and the VB-Eval dataset for the OSSI task. Moreover, we compile several subsets of VB2 to investigate the influence of data scale on model performance. The acoustic features are 80-dimensional log Mel-filterbank energies with a frame length of 25ms and a hop size of 10ms. The input frame length is fixed at 200 or 500 frames.

Model Usage. Our approach employs ResNet-based[23] models of various sizes, complemented by two pooling methods: Attentive Statistic Pooling (ASP)[24] and Temporal Statistic Pooling (TSP). To further harness the latent potential of data, we employ the Simple Attention Modules (SimAM) to extract

more discriminative speaker embeddings[25, 26]. Additionally, we introduce the ResNet50-based face recognition model for comparative analysis. This model is trained by Glint360K[27], which comprises 17,091,657 faces of 360,232 individuals.

Training strategies. We incorporate two different strategies for model training as follows:

- **Pre-train on the VC2 and fine-tune on the mixed set.** Following the findings in the [6], the Mix-FT strategy demonstrates the capability to further enhance the performance of speaker recognition systems.
- **Pre-train on the VB2 and fine-tune on the VC2 set.** Inspired by the findings of the LLM, models trained with massive data exhibit stronger generalization abilities. Moreover, by fine-tuning the VC2 set, the highly generalized model can learn more refined features.

Specifically, in both pre-training stages, we adopt the on-the-fly data augmentation for the variation of data and the speed perturbation to triple the number of speakers. The SGD optimizer updates the model parameters, and the StepLR scheduler with an initial learning rate of 0.1 decays to 1e-4 until convergence. For the fine-tuning stage, the Large-Margin Fine-Tune (LMFT) strategy[28] is introduced, accompanied by the removal of data augmentation. The LR in this phase must be set lower than the pre-training phase, and employing a relative smaller LR for a larger model has been found to be more effective.

5. Results

5.1. Speaker Verification

5.1.1. Different Strategies

We adopt the ResNet100-ASP with the simple attention module as the speaker encoder to generate speaker embeddings. In addition, we curate several randomly sampled subsets of the VB2 dataset, each containing varying speaker counts: 5k, 10k, 30k, and the full version with 110k speakers. As shown in Tab.3, increasing the number of mixed fine-tuned speakers does not consistently lead to significant performance improvements. Besides, compared to the Mix-FT mentioned previously, Fine-tuning the model on a large-scale pre-trained dataset results in a notable 43.4% in EER reduction on the VoxCeleb1-O set. Since this strategy is more intuitive and effective, the following experiments follow this training pattern.

5.1.2. Different Pre-train Data Scales

For a more detailed examination of performance variations with changes in data volume, we randomly compile diverse sub-sets of the VB2 set with different size of speakers. As illustrated in Figure 3, an increase in the number of speakers correlates with enhanced performance. It can be indicated that by stacking more data during the pre-training phase, the model becomes

Table 3: The system performance on the VoxCeleb1-O test set based on different training data for different stages. The LMFT and other post-processing strategies have not been introduced.

Pre-train	Fine-tune	EER[%]	minDCF _{0.01}
VC2	-	0.606	0.052
VC2	VC2+VB2-5K	0.527	0.047
VC2	VC2+VB2-10K	0.505	0.049
VC2	VC2+VB2-30K	0.505	0.051
VC2	VC2+VB2	0.674	0.066
VB2	-	0.893	0.093
VB2	VC2	0.340	0.026

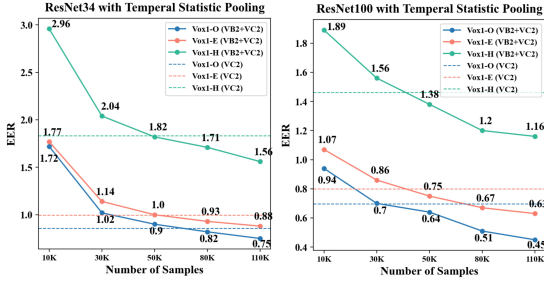


Figure 3: The EER and minDCF performance on VoxCeleb1-test set. ResNet34-TSP and ResNet100-TSP models, trained with different data scales, are pre-trained and then fine-tuned on VC2. Dotted lines represents system performances directly trained on VC2 for comparison. LMFT and other post-processing strategies are not included in this analysis.

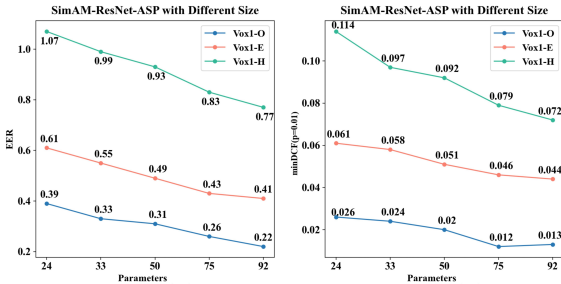


Figure 4: The EER and minDCF performances on VoxCeleb1-test set. All the ResNet-based models (ResNet34, 50, 100, 221, 293) embedded with the SimAM and the ASP, are pre-trained on the full VB2 set and then fine-tuned on VC2. The LMFT strategy is introduced during the fine-tuning stage.

more robust and generalized to adapt to diverse domains. From another perspective, in line with the scaling law principle, amplifying the data volume requires enlarging the model to achieve more significant effects, which is also illustrated in Fig.3. When we increase the number of model parameters, the decreases in EER on all test sets become steeper.

5.1.3. Different Model Complexity

To explore the performance bounds brought about by the increase in data volumes, we progressively escalate the model complexity. As shown in Fig.4, we observe continuous boosting performance on the Vox1-test set with the model size expansion.

Furthermore, we adopt the same settings as [29] for post-processing on the ResNet293-based model and finally achieve state-of-the-art performance. As shown in Tab.4, the EER and the minDCF can be reduced to 0.17% and 0.006% on the VoxCeleb1-O test set, respectively.

5.2. Open-Set Speaker-Identification

To assess the influence of data scale and modality on the OSSSI, we adopt the ResNet50 as the backbone to train the feature extractor and utilize the DIR at different FARs to evaluate the OSSSI performance. As depicted in Tab.5, increasing the num-

Table 4: The post-processing results based on the SimAM-ResNet293 single system. The p_{target} is set 0.01.

Method	Vox1-O		Vox1-E		Vox1-H	
	EER	mDCF	EER	mDCF	EER	mDCF
ResNet293	0.23	0.013	0.42	0.044	0.77	0.072
+AS-Norm	0.22	0.009	0.40	0.042	0.73	0.073
++QMF	0.17	0.006	0.37	0.037	0.68	0.070

Table 5: The baseline of the OSSSI based on ResNet50 (Pre-trained on VB2, holding a 1.02% EER on VoxCeleb-O test set), reflecting the DIR performance at different FAR. Enroll nums means the number of utterances included in the gallery set per speaker. The DIR@FAR=1 denotes there is no rejection.

Protocol Type	Enroll nums	DIR@FAR [%]			
		0.001	0.01	0.1	1
VB-Eval-S	1	88.10	93.35	93.46	98.08
	3	96.32	97.98	98.74	99.02
	5	96.97	98.04	98.68	99.11
VB-Eval-M	1	68.80	82.95	90.40	94.36
	3	86.86	92.37	95.95	97.69
	5	91.60	94.90	96.91	98.22
VB-Eval-L	1	20.73	66.37	80.95	88.09
	3	23.17	83.65	92.66	95.86
	5	24.94	87.12	94.34	96.72

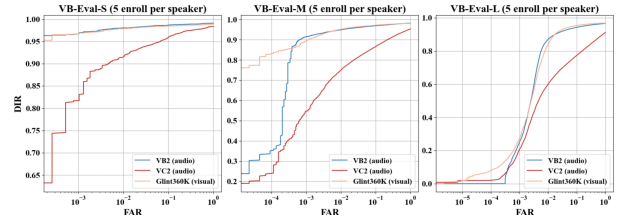


Figure 5: The ROC curve of two speaker recognition models and a face recognition model.

ber of enrollment utterances exhibits a positive association with the enhancement of DIR@FAR. However, as the gallery size expands exponentially, a noticeable decline in performance is observed, indicating the need for further studies.

In another modality, we utilize the same backbone pre-trained with Glint360K as the face recognition model, achieving a remarkable 0.03% EER on the VoxCeleb1-O test set[30]. Although the training dataset of speaker recognition differs from that of face recognition, the scales of data are comparable (110K vs 360K). Subsequently, we extract faces from videos in both the gallery and probe sets at intervals of 0.3 seconds. By encoding the faces from the same video and then averaging the face embeddings, we obtain the utterance-level face embedding.

As shown in Fig.5, we observe that large-scale training of facial and speaker recognition models yields relatively comparable results in both VB-Eval-S and VB-Eval-L. While the models show outstanding results in the VB-Eval-S, they reflect degraded performance in the VB-Eval-L. We speculate that the key factor is the insufficient training data for both modalities. Additionally, there is a slight discrepancy in performance between the speaker and face models in VB-Eval-M. However, the distance between the speaker model and the face model for recognition can be narrowed by enlarging the quantity of data.

6. Conclusion

This paper provides a large-scale audio-visual corpus for speaker recognition, comprising over 110K individuals gathered from YouTube. Through a series of ablation studies, we investigate the impact of training strategies, data scale, and model complexity on speaker verification, achieving state-of-the-art results. In addition, we introduce a new Open-Set Speaker-Identification benchmark alongside relevant baseline metrics derived from the VoxBlink-clean dataset. Notably, our findings reveal that speaker recognition models trained on comparable data scales and utilizing similar architectures as facial recognition models demonstrate comparable performance.

7. Acknowledgement

This research is funded by China Mobile Research Institute-Wuhan University Systematic Artificial Intelligence Collaboration Platform under the project "Joint Research and Development of Large Voiceprint Model for Telephone Scenarios". Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

8. References

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc of CVPR*, 2016, pp. 770–778.
- [2] J. Thienpondt and K. Demuynck, "Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings," in *Proc of ASRU*, 2023, pp. 1–8.
- [3] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking," in *Proc of INTERSPEECH*, 2023, pp. 5301–5305.
- [4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc of INTERSPEECH*, 2020, pp. 3830–3834.
- [5] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. Meng, "MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification," in *Proc of INTERSPEECH*, 2022, pp. 306–310.
- [6] X. Qin, D. Cai, and M. Li, "Robust multi-channel far-field speaker verification under different in-domain data availability scenarios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 71–85, 2023.
- [7] X. Qin, N. Li, W. Chao, D. Su, and M. Li, "Cross-Age Speaker Verification: Learning Age-Invariant Speaker Embeddings," in *Proc of INTERSPEECH*, 2022, pp. 1436–1440.
- [8] B. Han, Z. Chen, Z. Zhou, and Y. Qian, "The SJTU System for Short-Duration Speaker Verification Challenge 2021," in *Proc of INTERSPEECH*, 2021, pp. 2332–2336.
- [9] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *Proc of ICASSP*, 2019, pp. 5816–5820.
- [10] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc of INTERSPEECH*, 2017.
- [11] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc of INTERSPEECH*, 2018, pp. 1086–1090.
- [12] N. Torgashov, R. Makarov, I. Yakovlev, P. Malov, A. Balykin, and A. Okhotnikov, "The id r&d voxceleb speaker recognition challenge 2023 system description," *arXiv preprint arXiv:2308.08294*, 2023.
- [13] I. Yakovlev, A. Okhotnikov, N. Torgashov, R. Makarov, Y. Vovodin, and K. Simonchik, "VoxTube: a multilingual speaker recognition dataset," in *Proc of INTERSPEECH*, 2023, pp. 2238–2242.
- [14] Y. Lin, X. Qin, G. Zhao, M. Cheng, N. Jiang, H. Wu, and M. Li, "Voxblink: A large scale speaker verification dataset on camera," *arXiv preprint arXiv:2308.07056*, 2023.
- [15] A. K. Jain and S. Z. Li, *Handbook of face recognition*. Springer, 2011, vol. 1.
- [16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [17] M. Cheng, H. Wang, Z. Wang, Q. Fu, and M. Li, "The whulibaba audio-visual speaker diarization system for the misp 2022 challenge," in *Proc of ICASSP*, 2023, pp. 1–2.
- [18] M. Cheng, W. Wang, X. Qin, Y. Lin, N. Jiang, G. Zhao, and M. Li, "The dku-msxf diarization system for the voxceleb speaker recognition challenge 2023," *arXiv preprint arXiv:2308.07595*, 2023.
- [19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc of CVPR*, 2019, pp. 4685–4694.
- [20] R. Peri, S. O. Sadjadi, and D. Garcia-Romero, "Voxwatch: An open-set speaker recognition benchmark on voxceleb," *arXiv preprint arXiv:2307.00169*, 2023.
- [21] A. Malegaonkar and A. Ariyaecinia, "Performance evaluation in open-set speaker identification," in *Biometrics and ID Management: COST 2101 European Workshop, BioID 2011, Brandenburg (Havel), Germany, March 8-10, 2011. Proceedings 3*. Springer, 2011, pp. 106–112.
- [22] M. Gunther, S. Cruz, E. M. Rudd, and T. E. Boulton, "Toward open-set face recognition," in *Proc of CVPR Workshops*, 2017, pp. 71–80.
- [23] W. Cai, J. Chen, and M. Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Proc of Odyssey*, 2018, pp. 74–81.
- [24] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc of INTERSPEECH*, 2018, pp. 2252–2256.
- [25] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *ICML*. PMLR, 2021, pp. 11 863–11 874.
- [26] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Simple attention module based speaker verification with iterative noisy label detection," in *Proc of ICASSP*, 2022, pp. 6722–6726.
- [27] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang *et al.*, "Partial fc: Training 10 million identities on a single machine," in *Proc of CVPR*, 2021, pp. 1445–1449.
- [28] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *Proc of ICASSP*, 2021, pp. 5814–5818.
- [29] Z. Li, Y. Lin, X. Qin, N. Jiang, G. Zhao, and M. Li, "The dku-msxf speaker verification system for the voxceleb speaker recognition challenge 2023," *arXiv preprint arXiv:2308.08766*, 2023.
- [30] R. Tao, K. A. Lee, Z. Shi, and H. Li, "Speaker recognition with two-step multi-modal deep cleansing," in *Proc of ICASSP*, 2023, pp. 1–5.