VOXBLINK: A LARGE SCALE SPEAKER VERIFICATION DATASET ON CAMERA

Yuke Lin[1,2]*,Xiaoyi Qin*[1,2]*, Guoqing Zhao*[3]*,Ming Cheng*[1,2]*, Ning Jiang* [3]*,Haiying Wu*[3]*, Ming Li*[1,2]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Duke Kunshan University, Kunshan, China
[3]Mashang Consumer Finance Co., Ltd

ABSTRACT

In this paper, we introduce a large-scale and high-quality audio-visual speaker verification dataset, named **VoxBlink**. We propose an innovative and robust automatic audio-visual data mining pipeline to curate this dataset, which contains 1.45M utterances from 38K speakers. Due to the inherent nature of automated data collection, introducing noisy data is inevitable. Therefore, we also utilize a multi-modal purification step to generate a cleaner version of the VoxBlink, named VoxBlink-clean, comprising 18K identities and 1.02M utterances. In contrast to the VoxCeleb, the VoxBlink sources from short videos of ordinary users, and the covered scenarios can better align with real-life situations. To our best knowledge, the VoxBlink dataset is one of the largest publicly available speaker verification datasets. Leveraging the VoxCeleb and VoxBlink-clean datasets together, we employ diverse speaker verification models with multiple architectural backbones to conduct comprehensive evaluations on the VoxCeleb test sets. Experimental results indicate a substantial enhancement in performance—ranging from 12% to 30% relatively—across various backbone architectures upon incorporating the VoxBlink-clean into the training process. The details of the dataset can be found on Site.

Index Terms— Speaker Verification, Dataset, Large-scale, Multi-modal.

1. INTRODUCTION

Automatic Speaker Verification (ASV) in wild scenarios has achieved remarkable success consisting of the evolution of backbone architecture [1, 2, 3], the introduction of diverse loss functions [4, 5], and the availability of large-scale corpora [6, 7]. Even though growing efforts have been devoted to refining networks and training strategies [8, 9], the academic community still faces constraints by the limited scale and diversity of available datasets.

In the computer vision field, millions of images establish a robust foundation for face recognition. Regrettably, in the field of speaker recognition, the availability of publicly accessible datasets with both millions of utterances and tens of thousands of speakers in the wild remains noticeably limited. As is shown in Fig 1, many contributions have been made to enrich ASV datasets [6, 7, 10, 11, 12, 13, 14, 15]. Some datasets [11, 12] mainly comprise multi-channel far-field speeches, which includes limited number of speakers. While others [10, 13, 14] fall short of their limited styles, languages and scales. Among these endeavors, the VoxCeleb [6, 7] stands out as the most successful as it contains over one million utterances from thousands of speakers. Nonetheless, compared to face recognition, the quantity remains relatively small. Recently, the VoxTube [15] dataset releases over 4M utterances for 5,040 speakers, making it one of the largest open-source speaker recognition datasets to date. However, due to its reliance solely on audio information for clustering, the accuracy of its derived labels may not be very convincing. Meanwhile, hard samples can be easily discarded during the filtration.

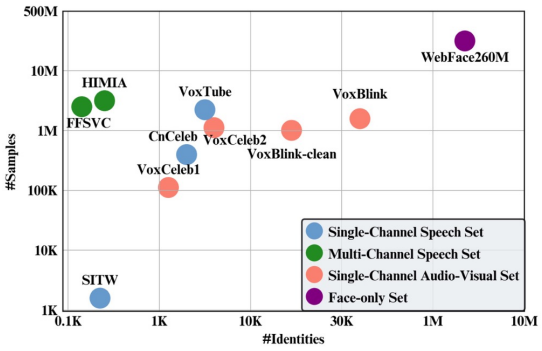

Fig. 1. Comparisons of # identities and # samples for our VoxBlink data and public ASV training set (also a face dataset as a contrast). The x and y-axis have **non-uniform** scales.

Therefore, we introduce a new large-scale audio-visual dataset for speaker verification, VoxBlink. All VoxBlink data is captured automatically from users who upload *short* videos on the YouTube platform, which contain over **1.4M utterances** from over **38K speakers**. In order to further purify the data without filtering out difficult samples, we use a multi-modal validation approach that results in a purified version of the VoxBlink (VoxBlink-clean), which contains 18k individuals and 1.02M utterances. All raw data is processed by multi-stage audio-visual models (including but not limited to face and lip detection, face verification, active speaker detection and overlap detection). Furthermore, by implementing audio-visual models for data mining, our data automatic pipeline exhibits greater resilience and promise in data size. Since data from the VoxBlink are mainly collected from the "wild", it also inherently exhibits diverse-

Corresponding Author: Ming Li.
This research is funded in part by the National Natural Science Foundation of China (62171207), Science and Technology Program of Suzhou City(SYC2022051) and MaShang Consumer Finance Co.Ltd. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

age, diverse-lingual, diverse-style and diverse-device attributes. As the VoxBlink is an audio-visual dataset, it can be used in various other applications, such as speech separation[16, 17], multi-modal verification[18, 19], and speaker diarization[20, 21], among others.

In addition, we also incorporate the VoxBlink-clean dataset in training various models with different backbones. When introducing the VoxBlink-clean, all models exhibit a performance boost ranging from 12% to 30% in terms of relative EER reduction on the VoxCeleb1-O test set. Our primary contributions can be summarized as follows: 1) We propose a more scalable and robust pipeline for mining speaker verification data; 2) We collect a large-scale audio-visual speaker verification dataset VoxBlink and its purified version VoxBlink-clean; 3) We achieve significant performance improvements under different backbones by integrating the VoxBlink-clean dataset into training process.

2. DATA MINING

2.1. Data Description

The VoxBlink contains 1,455,190 utterances from 38,065 channels on YouTube, while its purified version VoxBlink-clean comprises 1,028,095 utterances from 18,381 speakers. All speech/video segments are extracted from short videos uploaded by ordinary YouTube users, encompassing various contexts, including podcasts, music lives, speeches, live streaming highlights, etc. Indoor reverberation, non-verbal voice, background music and other acoustic conditions have increased the complexity and diversity of the data. Most of the segments are recorded on mobile devices, with recording environments spanning indoors, outdoors, and a variety of complex scenarios. Other statistic information can be found in Table 1 and Fig 3 shows a visualization of the statistics. The majority of speakers within the VoxBlink dataset are female, and its purified version is relatively gender-balanced. The dataset is multi-lingual yet English-dominant, with participants ranging from over 130 different regions worldwide.

Table 1. *Statistics for the VoxBlink dataset. The last two rows describe the* ***time-varying*** *characteristics across videos recorded by the same speaker.*

Dataset	VoxBlink	VoxBlink-clean
# of SPKs	38,065	18,381
# of male-SPKs	15,013	8,124
# of videos	372,084	241,170
# of hours	2,135	1,670
# of utterances	1,455,190	1,028,095
Avg # of videos per SPK	9.77	13.12
Avg # of utterances per SPK	38.23	55.93
Avg # of duration per utterance (s)	5.28	4.87
Avg # of video recording intervals (days)	39.72	34.55
Avg # of video recording span (days)	440.07	441.85

2.2. Collection Pipeline

As depicted in Fig 2, we employ an automatic multi-modal data-mining pipeline to construct the VoxBlink database from YouTube. The novelty of our approach lies in utilizing user avatars for

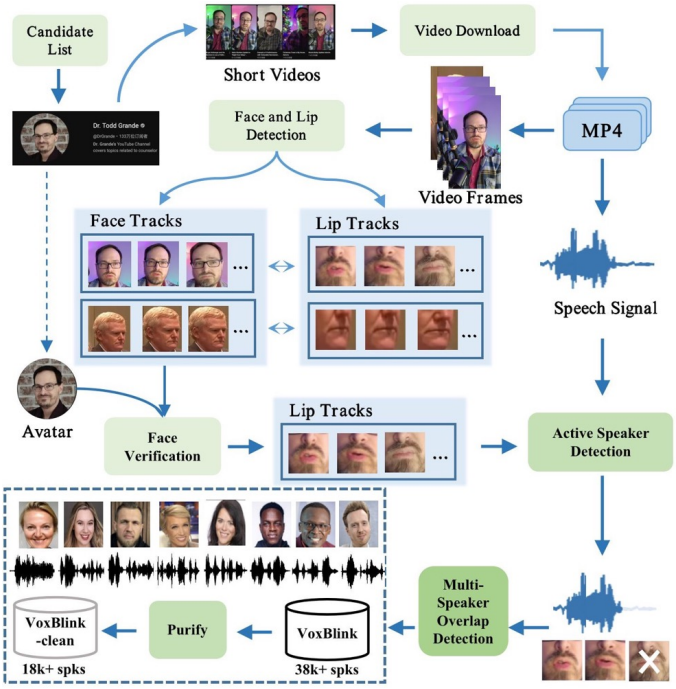

Fig. 2. The automatic pipeline for the VoxBlink dataset.

frame-by-frame face verification and lip motions for active speaker detection. Additionally, with the help of other auxiliary tools, we can extract speech/video segments specifically pertaining to the target user. For clarity, we summarize the processes as follows:

Step I: Candidate Collection. We start by compiling a list containing over 2,000 commonly used names as well as some professions and themes for data diversity. Having observed that users are more likely to appear in short videos, we opt to retain users with a single-face avatar who have uploaded short videos. Over 1M videos from 61,038 users with avatars are downloaded after duplicate removal in the YouTube retrieval. Since the data source of the VoxBlink consists of only short videos, there should be no overlapping with the VoxCeleb and VoxMovies[22] datasets.

Step II: Face and lip tracking. Using the Retina Face[23] model, we detect both face and lip movements, producing corresponding video and lip tracks. By setting a threshold for the minimum Intersection Over Union (IOU) value between two consecutive detections, we ensure that each track contains only one face or lip sequence.

Step III: Face verification. After face and lip tracking, we utilize the ResNet-IRSE50 [4] model to extract face embeddings for each speaker frame by frame along each track. Meanwhile, the face embedding of the avatar has been extracted as template embedding using flip augmentation, which promotes template robustness. Then, cosine similarities are calculated along the track, and the track-level average score is calculated to discard non-target tracks.

Step IV: Active Speaker Detection. To refine the track and eliminate silent or out-of-sync segments, we utilize a Seq2Seq audio-visual speaker diarisation model [20]. This model leverages lip motions and audio cues to identify instances of active speech within the track. This approach not only facilitates the removal of

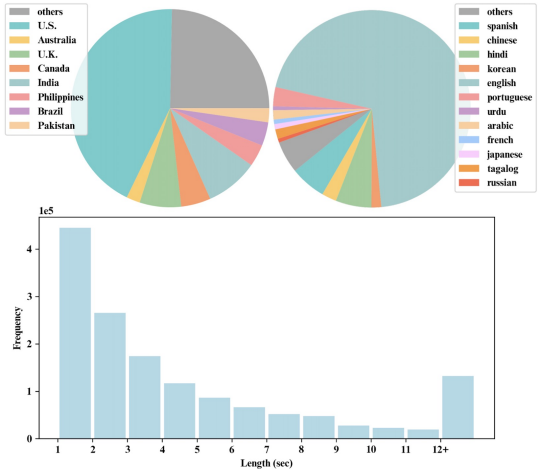

Fig. 3. **Top:** the distribution of geographic locations (*left*) and languages (*right*) of speakers. **Bottom**: The distribution of utterance lengths in the dataset.

silent or voice-over sections but also effectively excludes out-of-sync fragments.

Step V: Multi-Speaker Overlap Detection. In order to mitigate the potential disruption caused by overlapping speech data and enhance the quality of speech segments, we employ a conformer-based Overlapping Speech Detection (OSD) toolkit [24]. Furthermore, we discard utterances shorter than one second in duration.

Step VI: Meta Information Collection. Due to platform constraints, only the geographical locations of approximately 21,000 speakers are recorded. Given the challenges in obtaining gender labels, we binary-classify the speakers' genders using audio-visual data. We also infer the utterance-level language labels using the Whisper's[25] base model and obtain speaker-level labels through a voting mechanism. Other meta-informations about the video, including each video's release time, category, and tags, are collected for other potential applications. All collection meta will be released together with data.

Finally, The audio-visual data obtained through the pipeline constitutes the VoxBlink, whose name is inspired by the characteristics of short videos. The threshold within the collection pipeline is intentionally relaxed to facilitate the accumulation of a larger volume of data.

2.3. Multi-modal Purify

Due to automated data collection and relatively relaxed threshold settings, the introduction of noisy data is inevitably unavoidable. Upon manual inspection, we have found that some recordings are merely lip-syncing, indicating the need to purify our data further. Therefore, we purify the VoxBlink dataset utilizing the following metrics:

- The average score of within-speaker speech embedding cosine similarities derived by ResNet34 [2].
- The average score of within-speaker face embedding cosine similarities derived by [4].
- The average music-speech discrimination score of a speaker by [26].

We retain only those speakers with five or more utterances in order to uphold diversity within each speaker's data. Audio samples surpassing the aforementioned scores will be considered into the clean subset, VoxBlink-clean. Through randomly sampling observations, there are very few instances of noisy labels in this subset.

3. SPEAKER VERIFICATION MODEL TRAINING

In this section, we describe the experimental settings and implementation details of several speaker verification systems. We suggest a Mix-FineTune(Mix-FT) [27] training strategy to incorporate the VoxBlink into the training set.

3.1. Model Settings

ResNet-based model. ResNet-based speaker verification model achieved success in past years. Therefore, we conducted experiments using the state-of-the-art(SOTA) ResNet models for comparative analysis. Initially, we employed a standard ResNet34 [2] followed by a temporal statistic pooling layer as our baseline system. Then, to further tap into the latent potential of the data, we employed a larger ResNet model mounted with attention mechanisms – specifically, ResNet100 with Simple Attention Module (SimAM)[28] and frequency-wise Squeeze-Excitation (fwSE) [29] modules. The attentive statistics pooling (ASP) is employed to capture the importance of different frames.

TDNN-based model. ECAPA-TDNN [1] is currently the most popular and SOTA TDNN-series model for speaker verification. We conducted experiments using ECAPA-TDNN with 1024 channels to observe the performance of TDNN on a larger dataset.

3.2. Implement details

Data Usage. We conducted experiments using the VoxCeleb2 development set and the VoxBlink-clean set. 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 frames.

Data Augmentation. We adopt on-the-fly data augmentation[30] to add additive background noise or convolutional reverberation noise for the time-domain waveform. Also, we apply speaker augmentation with speed perturbation [31]. We speed up or down each utterance by a factor of 0.9 or 1.1, yielding shifted pitch utterances that are considered from new speakers. Finally, the training data contains 6,360,345 utterances (3,084,318 from the VoxBlink-clean and 3,276,027 from the VoxCeleb2) from 73,125 speakers (55,143 from the VoxBlink-clean and 17,982 from the VoxCeleb2).

Training Strategy. Inspired by [27], we observe that initiating training both datasets (VoxCeleb2 and VoxBlink-clean) from the start yields similar performance outcomes as beginning solely with the VoxCeleb2 and later incorporating the VoxBlink-clean for fine-tuning. Therefore, we carry most of our experiments based on three stages:

Stage.1 Warm-up. We perform a linear warm-up learning rate schedule at the first five epochs to the initial learning rate at 0.1, which aims to prevent model vibration and speed model training.

Table 2. *The experimental results of different backbones with/without the VoxBlink-clean dataset. All the benchmarks are based on the cosine scores between trials.* ***No post-processing*** *operations have been employed, such as LMFT, score norm, and QMF.* Δ *represents the relative EER reduction on the VoxCeleb1-O trials when using the VoxBlink-clean for Mix-FT compared to not using the VoxBlink-clean.*

ID	Model	Size	VoxCeleb2	VoxBlink -clean	Δ	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
						EER[%]	$\text{mDCF}_{0.01}$	EER[%]	$\text{mDCF}_{0.01}$	EER[%]	$\text{mDCF}_{0.01}$
M1	ResNet34-TSP	23.9M	×	✓	-	2.499	0.241	-	-	-	-
			✓	×	-	0.856	0.084	0.995	0.112	1.832	0.179
			✓	✓	**13.1%**	**0.744**	**0.057**	0.988	0.109	1.787	0.176
M2	ECAPA-TDNN	14.7M	✓	×	-	0.856	0.081	1.078	0.118	2.059	0.197
			✓	✓	**12.5%**	**0.749**	**0.077**	0.953	0.105	1.823	0.177
M3	SimAM-ResNet100-ASP	50.2M	✓	×	-	0.622	0.058	0.761	0.083	1.391	0.132
			✓	✓	**29.1%**	**0.441**	**0.044**	0.681	0.075	1.268	0.125
M4	fwSE-ResNet100-ASP	50.6M	✓	×	-	0.580	0.057	0.775	0.083	1.438	0.141
			✓	✓	**22.1%**	**0.452**	**0.038**	0.709	0.079	1.277	0.128

Stage.2 Plateau. The SGD optimizer updates the model parameters, and the StepLR scheduler with 0.1 initial LR drops to 1e-4 in 30 epochs. The step size is set to 10.

Stage.3 Mix-FT. As the first two stages only use the VoxCeleb2 dev set, we introduce the VoxBlink-clean in the last phase for the Mix-FineTuning(Mix-FT). The training process resumes at 1e-3 LR and gradually drops till convergence.

Finally, we adopt the Equal Error Rate (EER) and Minimum Detection Cost Function (minDCF) to measure system performance. Cosine similarity scores are calculated in the evaluation phase. As for the back end, we utilize the AAM-Softmax [4] (m=0.2, s=32) to classify different speakers.

4. EXPERIMENTIAL RESULTS

4.1. Base Results

As shown in table 2, the domain of the VoxBlink dataset does not closely align with that of the VoxCeleb2 dataset, as the results on the VoxCeleb1-O trials exhibit better performance when trained only on the VoxCeleb2. However, the VoxBlink-clean can be regarded as a supplementary training set of the VoxCeleb2. As we can see, across models M1 to M4, we achieve performance enhancements of relative 13.1%, 12.5%, 29.1%, and 22.1% when introducing the VoxBlink-clean, respectively. Performance improvements are observed across all other test protocols (VoxCeleb1-E and VoxCeleb1-H) as well, relatively ranging from 2% to 12%. Moreover, as we enlarge the model size, the positive impact of adding the VoxBlink-clean for training becomes increasingly noticeable.

4.2. LMFT and Score calibration

The ASV systems could benefit from several post-processing methods, including Large-Margin Fine-Tune (LMFT) [8], Adaptive Symmetric Score Normalization (AS-Norm) [32] and Quality Measure Functions (QMF) [8]. Therefore, we follow the same post-processing settings as [33] to enhance performance. As shown in Table 3, by incorporating the VoxBlink-clean set for Mit-FT training, followed by a series of post-processing steps, we achieved a reduction in EER from 0.441% to 0.282% on the Vox-O test set. Compared to using only the VoxCeleb2 as the training set with post-processing, we achieve a 20.8% relative EER reduction (0.356% to 0.282%).

Nevertheless, without incorporating the VoxBlink-clean for training, the LMFT achieves a 13.7% improvement, while the Mix-FT using the VoxBlink-clean shows only a 6.1% boost. We also observed that the EER reduction of the LMFT under the Mix-FT is not that significant in other models. We speculate that this might be because the average speech duration of the VoxBlink is shorter than that of the VoxCeleb2, meaning that less information is carried on for each utterance.

Table 3. *The post-processing results based on the SimAM-ResNet100 single system with/without the VoxBlink-clean data in training.*

ID	Method	Δ	VoxCeleb1-O	
			EER[%]	$\text{mDCF}_{0.01}$
Only VoxCeleb2 for training				
M3	SimAM-ResNet100	-	0.622	0.058
	+LMFT	13.7%	0.537	0.045
	++ AS-Norm	8.9%	0.489	0.047
	+++ QMF	27.2%	0.356	0.040
VoxCeleb2 for training and VoxBlink-clean for Mix-FT				
M3	SimAM-ResNet100	-	0.441	0.044
	+LMFT	6.1%	0.414	0.035
	++ AS-Norm	14.0%	0.356	0.037
	+++ QMF	21.8%	**0.282**	0.029

5. CONCLUSION

This paper introduces a large-scale audio-visual dataset named VoxBlink for the speaker verification task. We develop an automatic multi-modal data-mining pipeline to extract target users' audio-visual segments on YouTube and further conduct multi-modal detectors to build the VoxBlink-clean subset. We also achieve significant improvements by incorporating the VoxBlink-clean into model training across different backbones, which proves that the VoxBlink-clean is an excellent supplementary dataset for training speaker verification models.

6. REFERENCES

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[2] 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.Odyssey*, 2018, pp. 74–81.

[3] 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. Interspeech*, 2022, pp. 306–310.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4685–4694.

[5] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, July 2017.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[7] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[8] 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. ICASSP*, 2021, pp. 5814–5818.

[9] D. Cai, W. Wang, M. Li, R. Xia, and C. Huang, "Pretraining conformer with asr for speaker verification," in *Proc. ICASSP*, 2023, pp. 1–5.

[10] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: A challenging chinese speaker recognition dataset," in *Proc. ICASSP*, 2020, pp. 7604–7608.

[11] X. Qin, M. Li, H. Bu, S. Narayanan, and H. Li, "The 2022 far-field speaker verification challenge: Exploring domain mismatch and semi-supervised learning under the far-field scenarios," *arXiv preprint arXiv:2209.05273*, 2022.

[12] X. Qin, H. Bu, and M. Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *Proc. ICASSP*, 2020, pp. 7609–7613.

[13] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The Speakers in the Wild (SITW) Speaker Recognition Database," in *Proc. Interspeech*, 2016, pp. 818–822.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[15] I. Yakovlev, A. Okhotnikov, N. Torgashov, R. Makarov, Y. Voevodin, and K. Simonchik, "VoxTube: a multilingual speaker recognition dataset," in *Proc. Interspeech*, 2023, pp. 2238–2242.

[16] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, M. Gogate, K. Dashtipour, A. Hussain, and P. Bell, "Avse challenge: Audio-visual speech enhancement challenge," in *Proc. SLT*, 2023, pp. 465–471.

[17] J. Lin, X. Cai, H. Dinkel, J. Chen, Z. Yan, Y. Wang, J. Zhang, Z. Wu, Y. Wang, and H. Meng, "Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction," in *Proc. ICASSP*, 2023, pp. 1–5.

[18] M. Liu, K. A. Lee, L. Wang, H. Zhang, C. Zeng, and J. Dang, "Cross-modal audio-visual co-learning for text-independent speaker verification," in *Proc. ICASSP*, 2023, pp. 1–5.

[19] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, J. Hernandez-Cordero *et al.*, "The 2019 nist audio-visual speaker recognition evaluation." in *Proc. Odyssey*, 2020, pp. 259–265.

[20] M. Cheng, H. Wang, Z. Wang, Q. Fu, and M. Li, "The whu-alibaba audio-visual speaker diarization system for the misp 2022 challenge," in *Proc. ICASSP*, 2023, pp. 1–2.

[21] Z. Wang, S. Wu, H. Chen, M.-K. He, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. Siniscalchi, O. Scharenborg *et al.*, "The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[22] A. Brown, J. Huh, A. Nagrani, J. S. Chung, and A. Zisserman, "Playing a part: Speaker verification at the movies," in *Proc. ICASSP*, 2020.

[23] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proc.CVPR*, 2020, pp. 5202–5211.

[24] 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.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[26] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," *arXiv preprint arXiv:1703.01789*, 2017.

[27] 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.

[28] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Simple attention module based speaker verification with iterative noisy label detection," in *Proc. ICASSP*, 2022, pp. 6722–6726.

[29] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Information in 2D ResNets to Enhance Speaker Verification," in *Proc. Interspeech*, 2021, pp. 2302–2306.

[30] W. Cai, J. Chen, J. Zhang, and M. Li, "On-the-fly data loader and utterance-level aggregation for speaker and language recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1038–1051, 2020.

[31] W. Wang, D. Cai, X. Qin, and M. Li, "The dku-dukeece systems for voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2010.12731*, 2020.

[32] P. Matějka, O. Novotný, O. Plchot, L. Burget, M. Sánchez, and J. Černocký, "Analysis of Score Normalization in Multilingual Speaker Recognition," in *Proc. Interspeech*, 2017.

[33] 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.