VoxAging: Continuously Tracking Speaker Aging with a Large-Scale Longitudinal Dataset in English and Mandarin

Zhiqi Ai[1], Meixuan Bao[1], Zhiyong Chen[1], Zhi Yang[1], Xinnuo Li[2], Shugong Xu[3,]*

[1]Shanghai University, China
[2]New York University, USA
[3]Xi'an Jiaotong-Liverpool University, China

aizhiqi-work@shu.edu.cn, shugong.xu@xjtlu.edu.cn

Abstract

The performance of speaker verification systems is adversely affected by speaker aging. However, due to challenges in data collection, particularly the lack of sustained and large-scale longitudinal data for individuals, research on speaker aging remains difficult. In this paper, we present VoxAging, a large-scale longitudinal dataset collected from 293 speakers (226 English speakers and 67 Mandarin speakers) over several years, with the longest time span reaching 17 years (approximately 900 weeks). For each speaker, the data were recorded at weekly intervals. We studied the phenomenon of speaker aging and its effects on advanced speaker verification systems, analyzed individual speaker aging processes, and explored the impact of factors such as age group and gender on speaker aging research.

Index Terms: speaker verification, speaker aging, longitudinal dataset

1. Introduction

Speaker recognition (SR) and face recognition (FR) are widely used biometric technologies for identity authentication [1, 2]. However, both face challenges related to aging [3, 4, 5]. As people age, physiological changes in the face and vocal tract lead to gradual alterations in their features, negatively affecting the accuracy of SR and FR systems. In SR systems, the impact of aging is particularly significant [3, 4, 6, 7, 8, 9]. Aging affects the vocal cords and vocal tract, causing voiceprint features to deteriorate, which reduces the reliability of SR systems [10, 11]. Consequently, SR systems require more frequent updates to ID templates to maintain performance, as voiceprint features are highly sensitive to aging-related changes.

Early research on speaker aging was limited by scarce data and the capabilities of SR models. These studies primarily relied on traditional speech datasets with short time spans [11, 12, 13]. For instance, [12] used SEARP pitch analysis and observed that healthy individuals exhibited less tremor during vowel production, whereas elderly individuals displayed more pronounced tremors. Similarly, [11] and [13] employed models such as GMM-UBM and found that speaker aging negatively impacted SR system performance, suggesting that incorporating age-related factors could enhance accuracy.

Recent studies have increasingly focused on the impact of speaker aging using cross-age speaker datasets. Research on the TCDSA dataset [4, 6, 7] shows that verification scores decline as the time span increases, with short-term aging effects being relatively minor. Over time, genuine speaker scores decrease significantly, while impostor scores remain stable [6]. A
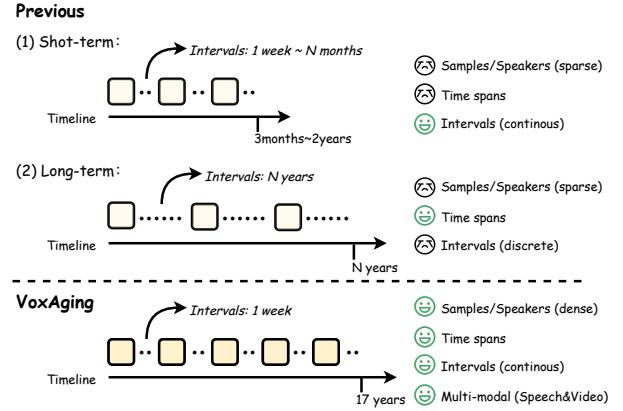

Figure 1: *Previous short-term datasets have continuous intervals but limited time spans, while long-term datasets have long time spans with discrete intervals, both with sparse sampling. The VoxAging offers dense sampling, continuous weekly intervals, long time spans, and multi-modal data.*

fixed decision threshold can exacerbate classification error rates even with just a few years' age difference [7]. Recent work on advanced SR models, such as ResNet34 and ECAPA-TDNN [8, 9], confirms that aging-related changes degrade system performance. These effects are more pronounced in female English speakers but have a greater impact on male Finnish speakers [9].

A major challenge in speaker aging research is the scarcity of long-term data. Most existing datasets cover relatively short periods (typically around 3 months to 2 years) with a limited number of speakers, leading to data jitter and outliers. For instance, the CSLT-Chronos dataset [14] includes 60 speakers and 84,000 samples collected over two years. Long-term datasets, such as TCDSA, contain recordings from 17 speakers over a span of 28 to 58 years but with fewer than 10 samples per speaker [4]. The LCFSH Finnish dataset [9] only covers two discrete intervals (20 and 40 years), while the VoxCeleb dataset [8, 9], annotated with age-related face models for aging analysis, still lacks sufficiently dense audio evaluation data for each speaker.

To address the challenges of speaker aging in SR systems, we present VoxAging, a large-scale longitudinal dataset. It includes recordings from 293 speakers (226 English and 67 Mandarin) over a span of 17 years, totaling 7,522 hours, with weekly samples. Our research investigates how aging affects voice features and the performance of advanced SR models, as well as the impact of age group and gender on speaker aging.

Table 1: *Comparison of existing speaker aging datasets. '-' indicates unavailable information. "Discrete" means datasets with long session intervals, where each ID has only a few samples. "Continuous" means datasets with short session intervals and continuous collection. "gradient*" indicates that the session intervals gradually increase over time.*

Dataset	# of Spks	# of Segments	# of Hours	# Max Span (years)	# Session Intervals	Language	Modality
Discrete							
TCDSA [4]	17	231	30	58	1∼23 years	English	Speech
LCFSH [9]	109	15,474	-	40	20 years	Finnish	Speech
VoxCeleb-AE [9]	670	79,063	<352	10	-	English	Speech
VoxCeleb-CA [8]	971	92,635	<352	20	-	English	Speech
Continous							
MARP [15]	60	-	-	3	2 months	English	Speech
CSLT-Chronos [14]	60	84,000	70	2	gradient*	Mandarin	Speech
SMIIP-TV [16]	373	325,049	305	0.25	4 days	Mandarin	Speech
VoxAging (Ours)	293	2,629,100	7,522	17	1 week	English, Mandarin	Speech, Video

2. VoxAging Dataset

2.1. Previous speaker aging datasets

As shown in Table 1, existing speaker aging datasets can be classified into two types: discrete and continuous, based on session intervals. Discrete datasets [4, 9, 8] have long session intervals and limited samples per speaker, spanning several years to two decades. For instance, TCDSA [4] includes recordings from 17 speakers over a span of 28 to 58 years, but with fewer than 10 samples per speaker. LCFSH [9], a Finnish dataset, has only two time spans: 20 and 40 years. VoxCeleb-AE [9] and VoxCeleb-CA [8], derived from the VoxCeleb [17] dataset (originally designed for general speaker recognition), feature imprecise age labels and limited samples per speaker (an average of 123 utterances).

In contrast, continuous datasets [15, 14, 16] feature shorter session intervals and higher collection frequencies, ranging from a few months to days. MARP [15] covers 60 speakers with a 2-month interval, CSLT-Chronos [14] includes 60 speakers over 2 years, with 14 sessions collected at gradient intervals, and SMIIP-TV [16], a recently collected dataset, tracks data from 373 individuals continuously over 3 months at a high cost.

2.2. Data description

The VoxAging dataset is a large-scale, longitudinal collection compiled from 293 speakers, including 226 English speakers (112 female, 114 male) and 67 Mandarin speakers (23 female, 44 male). The dataset spans up to 17 years (approximately 900 weeks) with weekly recordings, offering dense sampling over an extended period. It contains 2,629,100 segments, amounting to 7,522 hours of audio-visual data. The data was sourced from YouTube[1] and Bilibili[2], with channels manually filtered to ensure high-quality videos and appropriate time spans. As shown in Table 1, the unique advantage of the VoxAging dataset lies in its continuous weekly intervals over such an extended period, setting it apart from previous speaker aging datasets, which have limited time spans or discrete intervals.

Figure 2 illustrates the static distribution of the VoxAging dataset. The time span and data size for English speakers are larger, primarily because Mandarin data collection is more challenging, with recordings often starting later (mostly after 2017).
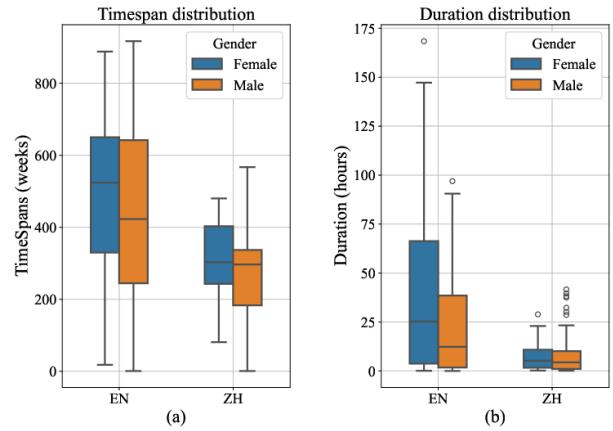

Figure 2: *VoxAging dataset distribution: (a) timespan distribution, (b) duration distribution. In VoxAging, there are 293 speakers: 226 English speakers (112 female and 114 male) and 67 Mandarin speakers (23 female and 44 male).*

For more detailed statistics, refer to the project page[3].

2.3. Collection pipeline

Our data cleaning process differs from traditional methods [17, 18] that rely on a single static template, as we place greater emphasis on the impact of individual aging on facial and voice features. To address this, we employ dynamic templates in the cleaning process to account for the aging of facial appearance and voice characteristics, as illustrated in Figure 3. The entire cleaning process is divided into three steps:

- **Step 1.** Video split via multi-modal methods. We segment videos into clips using multi-modal methods, including shot boundary detection[4] to identify scene transitions, YOLO-world [19] for person detection, and voice activity detection [20] to isolate speech segments. The intersection of visual and audio boundaries is then calculated to define each segment.

- **Step 2.** Longitudinal data cleaning with dynamic templates. We employ dynamic templates for data cleaning and use face

---

[1] https://www.youtube.com
[2] https://www.bilibili.com

[3] https://github.com/aizhiqi-work/voxaging
[4] https://www.scenedetect.com

recognition [21] and speaker verification [22] models to extract feature representations for each segment. Then, we apply the DBSCAN clustering algorithm to group similar speaker identities from different periods, removing noisy data and ensuring ID consistency. Finally, these dynamic templates are used to refine the cleaning process for each time segment.

- **Step 3.** Multi-experts labeling & noise reduction. We utilize multiple expert models to annotate and refine the cleaned data. Specifically, we employ a speech transcription model [23], a multi-modal emotion recognition model [24, 25], and an age estimation model [24] to label the data. The age estimation model is particularly crucial, as it assigns age groups to each ID. During the initial data collection, we could only determine the timespan of each video, without knowing the user's actual age. Finally, we apply speech enhancement models [26] to the high-quality data for noise reduction, further improving the accuracy of age analysis.

3. Experiments

3.1. Data setting

As shown in Table 2, the data settings of VoxAging include "X-Independent" and "X-Dependent" configurations.

- The "X-Independent" setting consists of two subsets: VoxAging-EN (English speakers) and VoxAging-ZH (Mandarin speakers). This setup investigates the impact of aging on speaker verification systems. VoxAging-EN is divided into 11 time spans (0 to 10 years), while VoxAging-ZH is divided into 5 time spans (0 to 4 years).

- The "X-Dependent" setting, using VoxAging-EN, explores the effects of age group (VoxAging-AgeGroup) and gender (VoxAging-Gender) on speaker aging. The dataset is divided into 5 age groups. It also includes 114 male and 112 female speakers. Both analyses cover 6 time spans: 0, 2, 4, 6, 8, and 10 years.
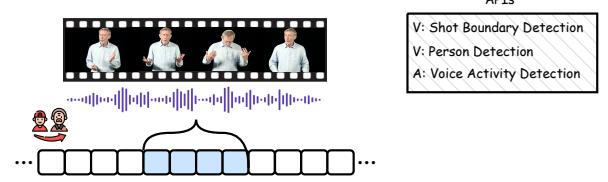
3.2. Model setting

As shown in Table 3, to investigate the impact of speaker aging on state-of-the-art speaker recognition models, we first employed the face recognition model ArcFace [21] as a baseline for aging. Subsequently, we evaluated seven advanced speaker recognition models [27], which demonstrated varying performances on the VoxCeleb dataset [17]. These models include RDINO [28], TDNN [29], SDPN [30], ECAPA-TDNN [22], CAM++ [31], ERes2Net [32], and ERes2Net-large [32]. Among these, the best-performing model was ERes2Net-large[5], achieving an EER of 0.57% on Vox-O[6].
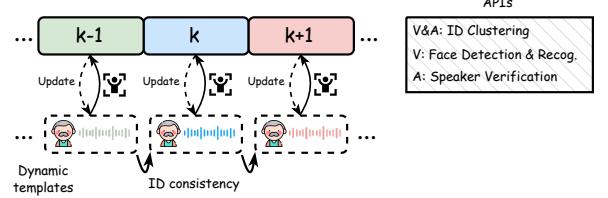
4. Results

4.1. Impact of speaker aging on advanced speaker verification systems

Table 3 shows the impact of speaker aging on advanced speaker verification systems. These models perform differently on the general test set Vox-O [17], and we use Equal Error Rate (EER) to evaluate the effect of aging on the VoxAging-EN and VoxAging-ZH subsets. As the time span increases, the EER of

---

[5] https://github.com/modelscope/3D-Speaker
[6] https://www.modelscope.cn/models/iic/speech_eres2net_large_sv_en_voxceleb_16k

Figure 3: *Illustration of the collection pipeline.*

Table 2: *Data setting for VoxAging.*

Setting		# of Spks	# of Trails
X-Independent			
VoxAging-EN	Cross-Age	226	1.1M
VoxAging-ZH	Cross-Age	67	0.5M
X-Dependent (EN)			
VoxAging-AgeGroup	<30	92	
	30~40	77	
	40~50	31	3.0M
	50~60	16	
	>60	10	
VoxAging-Gender	Male	114	1.2M
	Famale	112	

the speaker verification system deteriorates, indicating that the speaker recognition accuracy declines over time. Additionally, we use the face recognition model (ArcFace [21]) as the baseline for aging analysis. Compared to the speaker verification model, ArcFace demonstrates greater robustness to facial aging, delivering exceptional performance. However, despite this robustness, recognition accuracy still declines over time, with the EER rising from 0.31% to 1.52%.

In VoxAging-EN, RDINO [28] and TDNN [29] show relatively poor performance, as reflected by their higher initial EERs and deterioration rates of 2.98% and 1.93%, respectively. In contrast, ECAPA-TDNN [22], ERes2Net [32], CAM++ [31], and ERes2Net-Large [32] exhibit lower initial EERs and slower deterioration rates, suggesting that improving the performance of speaker recognition models can enhance their robustness against speaker aging. In VoxAging-ZH, all models display generally higher initial EERs and greater deterioration rates (significantly higher than in VoxAging-EN), but the overall trend remains consistent with VoxAging-EN.

Table 3: *Impact of speaker aging on advanced speaker verification systems.*

Model	Vox-O	VoxAging-EN EER(%)↓												VoxAging-ZH EER(%)↓					
		0	1	2	3	4	5	6	7	8	9	10	Δ	0	1	2	3	4	Δ
Face Modality																			
ArcFace [21]	-	0.31	0.38	0.55	0.72	0.75	1.00	1.23	1.23	1.42	1.56	1.52	1.21	0.52	0.58	0.62	0.75	0.82	0.30
Speech Modality																			
RDINO [28]	3.16	6.19	6.42	6.82	7.25	7.27	7.61	7.92	8.09	8.51	8.72	9.17	2.98	17.70	19.31	19.68	20.32	20.16	2.46
TDNN [29]	2.22	6.43	6.59	6.85	7.15	7.15	7.37	7.53	7.65	7.81	8.23	8.36	1.93	**9.06**	**9.82**	**10.63**	**11.58**	**11.78**	2.72
SDPN [30]	1.88	**2.86**	**2.91**	**3.00**	**3.20**	**3.18**	**3.20**	**3.31**	**3.45**	**3.55**	**3.78**	**3.73**	**0.87**	13.98	15.76	16.50	17.15	16.34	<u>2.36</u>
ECAPA-TDNN [22]	0.86	4.07	4.16	4.47	4.49	4.52	4.53	4.66	4.86	5.04	5.27	5.36	1.29	11.15	12.77	14.02	14.88	14.75	3.60
ERes2Net [32]	0.83	3.02	3.26	3.40	3.61	3.56	3.57	3.67	3.75	3.87	4.08	4.20	1.18	<u>10.30</u>	<u>11.08</u>	<u>12.18</u>	<u>12.57</u>	<u>12.43</u>	**2.13**
CAM++ [31]	<u>0.65</u>	3.72	3.94	4.13	4.19	4.31	4.19	4.29	4.46	4.64	4.76	4.80	1.08	12.53	14.37	15.91	16.44	16.37	3.84
ERes2Net-large [32]	**0.57**	<u>2.89</u>	<u>3.05</u>	<u>3.11</u>	<u>3.24</u>	<u>3.22</u>	<u>3.24</u>	<u>3.37</u>	<u>3.47</u>	<u>3.66</u>	<u>3.80</u>	<u>3.87</u>	<u>0.98</u>	10.52	11.87	12.91	13.74	13.72	3.20

However, there are some special cases. In VoxAging-EN, the initial EER and deterioration rate of SDPN [30] are comparable to those of ERes2Net-Large. In VoxAging-ZH, the initial EER of TDNN is relatively low, at only 9.06%.

4.2. Speaker similarity scores over time

Figure 4 shows the trend of speaker similarity scores over time in VoxAging, where embeddings were extracted using ECAPA-TDNN [22]. We randomly selected 10 English and 10 Mandarin speakers from the dataset and analyzed speaker similarity using a cubic polynomial fitting method over a weekly time span. The results show that speaker similarity decreases over time from the point of enrollment. This decline is caused by age-related changes in the speakers' voices, which emphasizes a key factor affecting the performance of speaker verification systems.

In Figure 4, the black dashed line represents the average trend of the speaker similarity score decline. It is clearly evident that there is a difference in the decay rate of speaker similarity between English and Mandarin. For the English average trend, it takes about 500 weeks ($\sim$10 years) for the speaker similarity to fall below the 0.5 threshold, while for the Mandarin average trend, it takes about 400 weeks ($\sim$8 years) for the speaker similarity to fall below the 0.5 threshold.

4.3. Impact of age group and gender on speaker aging

Table 4 shows the impact of age group and gender on speaker aging, with embeddings extracted using ERes2Net-Large [32]. In all age groups, the performance of the speaker verification system deteriorates with age. In VoxAging-AgeGroup, the initial EER for the young age group ($<$30 years) is relatively high, reaching 5.24% at the 10-year mark. The initial EER for the 30$\sim$40 and 40$\sim$50 age groups is lower than that of the young group, but the aging effect is more pronounced, with deterioration rates of 1.50% and 1.67%, respectively. The initial EER for the 50$\sim$60 age group is similar to that of the 40$\sim$50 group, but the deterioration is slower (1.17%). For those over 60 years old, the aging effect is the least pronounced, and the overall EER remains relatively stable, with a deterioration rate of 0.30%. Overall, the experiment shows that age-related voice changes are particularly significant in the 40$\sim$50 age group.

In VoxAging-Gender, it is clear that male speakers have lower initial EER values than female speakers. Additionally, both genders exhibit similar trends, with EER values increasing over time. The deterioration is more pronounced in the female group, with a deterioration rate of 2.62%, reaching an EER of 6.77% at the 10-year mark, higher than the male group (4.09%). This suggests that age-related voice changes may have a more noticeable impact on female group in VoxAging.
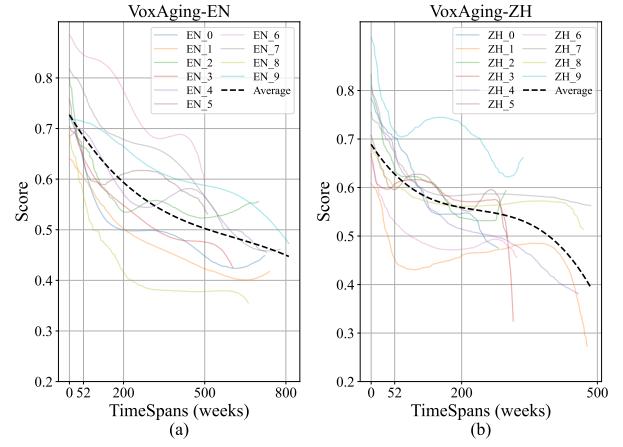

Figure 4: *Speaker similarity scores over time in VoxAging. Dashed black line indicates the average aging trend.*

Table 4: *The impact of age group and gender on speaker aging.*

Setting		EER(%)↓						
		0	2	4	6	8	10	Δ
AgeGroup	$<$30	4.12	4.58	4.64	4.76	5.14	5.24	1.12
	30$\sim$40	3.43	3.63	3.84	4.20	4.58	4.93	1.50
	40$\sim$50	2.57	2.70	3.10	3.26	3.62	4.24	**1.67**
	50$\sim$60	2.58	2.70	2.90	2.96	3.33	3.75	1.17
	$>$60	2.82	2.91	2.74	2.91	3.22	3.12	0.30
Gender	Male	3.54	3.71	3.77	3.75	3.92	4.09	0.55
	Female	4.15	4.77	5.05	5.50	6.02	6.77	**2.62**

5. Conclusions

In this paper, we present VoxAging, a large-scale longitudinal dataset. It includes recordings from 293 speakers (226 English and 67 Mandarin) over a span of 17 years, totaling 7,522 hours, with weekly samples. Our analysis of speaker aging reveals that the performance of speaker verification systems deteriorates with age. Improving the performance of speaker recognition models can enhance their resistance to speaker aging. Additionally, speaker similarity scores significantly declines over time. The impact of age and gender on speaker aging shows that 40$\sim$50 age group and female group exhibit more pronounced voice deterioration.

6. References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[2] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79 236–79 263, 2021.

[3] X. Qin, N. Li, S. Duan, and M. Li, "Investigating long-term and short-term time-varying speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3408–3423, 2024.

[4] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data," in *2012 5th IAPR international conference on biometrics (ICB)*. IEEE, 2012, pp. 478–483.

[5] K. Baruni, N. Mokoena, M. Veeraragoo, and R. Holder, "Age invariant face recognition methods: A review," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2021, pp. 1657–1662.

[6] F. Kelly and J. H. L. Hansen, "Evaluation and calibration of short-term aging effects in speaker verification," in *Interspeech 2015*, 2015, pp. 224–228.

[7] F. Kelly and J. H. Hansen, "Score-aging calibration for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2414–2424, 2016.

[8] X. Qin, N. Li, W. Chao, D. Su, and M. Li, "Cross-age speaker verification: Learning age-invariant speaker embeddings," in *Interspeech 2022*, 2022, pp. 1436–1440.

[9] V. P. Singh, M. Sahidullah, and T. Kinnunen, "Speaker verification across ages: Investigating deep speaker embedding sensitivity to age mismatch in enrollment and test speech," in *Interspeech 2023*, 2023, pp. 1948–1952.

[10] "Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers," *Speech Communication*, vol. 52, no. 7, pp. 638–651, 2010.

[11] Y. Lei and J. H. Hansen, "The role of age in factor analysis for speaker identification," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[12] L. A. Ramig and R. L. Ringel, "Effects of physiological aging on selected acoustic characteristics of voice," *Journal of Speech, Language, and Hearing Research*, vol. 26, no. 1, pp. 22–30, 1983.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[14] L. Wang, J. Wang, L. Li, T. F. Zheng, and F. K. Soong, "Improving speaker verification performance against long-term speaker variability," *Speech Communication*, vol. 79, pp. 14–29, 2016.

[15] A. D. Lawson, A. R. Stauffer, E. J. Cupples, S. J. Wenndt, W. P. Bray, and J. J. Grieco, "The multi-session audio research project (marp) corpus: goals, design and initial findings," in *Interspeech 2009*, 2009, pp. 1811–1814.

[16] X. Qin, N. Li, S. Duan, and M. Li, "Investigating long-term and short-term time-varying speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[17] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017*, 2017, pp. 2616–2620.

[18] L. Li, X. Li, H. Jiang, C. Chen, R. Hou, and D. Wang, "Cn-celeb-av: A multi-genre audio-visual dataset for person recognition," in *Interspeech 2023*, 2023, pp. 2118–2122.

[19] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 901–16 911.

[20] Z. Gao, Z. Li, J. Wang, H. Luo, X. Shi, M. Chen, Y. Li, L. Zuo, Z. Du, and S. Zhang, "Funasr: A fundamental end-to-end speech recognition toolkit," in *Interspeech 2023*, 2023, pp. 1593–1597.

[21] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[24] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4.

[25] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," in *Interspeech 2024*, 2024, pp. 1580–1584.

[26] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate what you describe: Language-queried audio source separation," *arXiv preprint arXiv:2203.15147*, 2022.

[27] S. Zheng, L. Cheng, Y. Chen, H. Wang, and Q. Chen, "3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement," *arXiv preprint arXiv:2306.15354*, 2023.

[28] Y. Chen, S. Zheng, H. Wang, L. Cheng, and Q. Chen, "Pushing the limits of self-supervised speaker verification using regularized distillation framework," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

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

[30] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, S. Zhang, and W. Wang, "Self-distillation prototypes network: Learning robust speaker representations without supervision," *arXiv preprint arXiv:2406.11169*, 2024.

[31] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332*, 2023.

[32] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, and J. Qi, "An enhanced res2net with local and global feature fusion for speaker verification," in *Interspeech 2023*, 2023, pp. 2228–2232.