



Investigating Long-Term and Short-Term Time-Varying Speaker Verification

Xiaoyi Qin, Na Li, Shufei Duan , and Ming Li , *Senior Member, IEEE*

Abstract—The performance of speaker verification systems can be adversely affected by time domain variations. However, limited research has been conducted on time-varying speaker verification due to the absence of appropriate datasets. This paper aims to investigate the impact of long-term and short-term time-varying in speaker verification and proposes solutions to mitigate these effects. For long-term speaker verification (i.e., cross-age speaker verification), we introduce an age-decoupling adversarial learning method to learn age-invariant speaker representation by mining age information from the VoxCeleb dataset. For short-term speaker verification, we collect the SMIP-TimeVarying (SMIP-TV) Dataset, which includes recordings at multiple time slots every day from 373 speakers for 90 consecutive days and other relevant meta information. Using this dataset, we analyze the time-varying of speaker embeddings and propose a novel but realistic time-varying speaker verification task, termed incremental sequence-pair speaker verification. This task involves continuous interaction between enrollment audios and a sequence of testing audios with the aim of improving performance over time. We introduce the template updating method to counter the negative effects over time, and then formulate the template updating processing as a Markov Decision Process and propose a template updating method based on deep reinforcement learning (DRL). The policy network of DRL is treated as an agent to determine if and how much should the template be updated. In summary, this paper releases our collected database, investigates both the long-term and short-term time-varying scenarios and provides insights and solutions into time-varying speaker verification.

Index Terms—Cross-age, reinforcement learning, speaker verification, template updating, time-varying.

I. INTRODUCTION

AUTOMATIC Speaker Verification (ASV) has made remarkable advancements in recent years, largely due to the

Manuscript received 27 July 2023; revised 21 January 2024 and 3 June 2024; accepted 6 July 2024. Date of publication 16 July 2024; date of current version 26 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62171207 and Grant 12004275, in part by the Science and Technology Program of Suzhou City under Grant SYC2022051, and in part by Tencent AI Lab Rhino-Bird Gift Fund. The associate editor coordinating the review of this article and approving it for publication was Dr. Omid Sadjadi. (*Corresponding author: Ming Li.*)

Xiaoyi Qin and Ming Li are with the School of Computer Science, Wuhan University, Wuhan 430072, China, and also with the Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Data Science Research Center, Duke Kunshan University, Suzhou 215316, China (e-mail: ming.li369@dukekunshan.edu.cn).

Na Li is with the Tencent AI Lab, Shenzhen 518054, China.

Shufei Duan is with the College of Electronic Information and Optical Engineering, Taiyuan University of Technology, Taiyuan 030024, China.

The source code and data resources are available on https://github.com/qinxiaoyi/TimeVarying_ASV.

Digital Object Identifier 10.1109/TASLP.2024.3428910

application of deep learning techniques such as X-vector [1] and its variant [2], [3], [4], which extract a fixed-dimensional discriminative feature from variable-length audio inputs. Margin-based loss functions such as SphereFace [5] and ArcFace [6], have also been adopted to train ASV system with large-scale databases, which effectively reduces the intra-speaker variability and increasing the inter-speaker distance. However, the current performance of ASV systems falls short of the standards required for certain applications. Several challenges and limitations exist that impede the reliability and accuracy of ASV in practical scenarios. We categorize these challenges into *speaker intrinsic* and *extrinsic variations*. *Speaker intrinsic variations* consist of variations in individual's speech and vocal characteristics, such as changes in the human voice due to aging, emotions, and physiological state. *Speaker external variabilities* also include various factors that can affect ASV performance. These factors include but not limited to low-quality data characterized by low signal-to-noise ratio (SNR), reverberation, distortion, and far-field recording. Additionally, cross-domain scenarios, such as cross-lingual and cross-channel conditions, can further degrade the ASV performance. While many research works have been devoted to speaker external factors [7], [8], [9], [10], [11], [12], limited attention has been given to *speaker intrinsic variations* due to challenges in simulating intrinsic changes and a scarcity of relevant data. However, in practical applications, voice characteristics of an individual can naturally vary over time, potentially leading to errors in speaker verification systems crossing a certain period of time [13], [14], [15], [16], [17]. Therefore, this paper focuses on the time-varying effects in speaker verification. We divide time-varying speaker verification into three subproblems:¹

- Intra-day variation speaker verification (IDV-SV) for variations across different times of the day;
- Short-term time-varying speaker verification (STTV-SV) for variations across times of the year;
- Long-term time-varying speaker verification or cross-age speaker verification (CA-SV) for variations across ages.

We focus on the solutions of STTV-SV and CA-SV. [18] introduces that the biometric template aging problem is typically addressed in the following ways: (1) frequent (and forced) template updates; (2) use of age invariant biometric features; (3) simulation of aging effects; (4) age progression compensation methods. Considering the time-varying speaker

¹The definition of time-varying scenarios in this paper is slightly different from [14], which defines the short-term as different times of the day, medium-term as times of the year, and long-term as changes with age.

characteristics of human speech signal, we adopt different strategies to handle different types of speaker time-varying challenges: proposing a template updating strategy to deal with short-term time-varying variabilities and learning age-invariant speaker representations to address long-term time-varying effects.

For STTV-SV scenarios, previous studies have shown that speaker verification systems can experience degraded performance over just a few months or even days [9], [16], [19]. However, existing short-term time-varying speaker verification datasets often have limited speakers and specific time intervals, and most of them are recorded in lab settings. Consequently, they may not fully capture the complexity of real-life scenarios and their proposed methods may lack robustness. Additionally, it is worth noting that some of these datasets were recorded decades ago, using a relatively low sampling rate of 8 kHz. To overcome these limitations and further investigate short-term time-varying speaker verification, we collect the SMIP-TimeVarying (SMIP-TV) dataset and publicly release it in this paper. The SMIP-TV dataset comprises continuous recordings of 373 speakers over a span of 90 days. This dataset also includes meta-information associated with each recording. In real-world usage scenarios, enrollment templates interact with positive or negative samples at various time periods. Therefore, we propose a novel time-varying speaker verification task called Incremental Sequence-pair Speaker Verification (ISSV), where enrollment template interacts with testing audios in chronological order, ASV system can continue updating templates in the interactive process of a sequential trial to counteract the effects of time-varying. In this task, we propose a template updating method that leverages deep reinforcement learning (DRL) to replace the fix-weight template updating approach [20], in which the updating thresholds and weights are pre-determined and fixed.

For CA-SV scenarios, also known as long-term time-varying speaker verification, early research focus on small-scale datasets due to the challenges in collecting cross-age speech data, which is time-consuming and expensive [21], [22], [23], [24]. Some recent works have analyzed the influence of age on speaker verification [25] and diarization [26]. However, the evaluation sets in these datasets do not include the cross-age speaker verification scenario. Some studies [27], [28], [29], [30], [31] also experimented on NIST SRE and TIMIT [32] datasets to estimate speaker age, but each speaker's data only cover one age point. [33] studied the impact of demographic imbalance on group fairness in speaker recognition, taking into account age influence. [34] establishes a quantitative measure between aging and ASV scores in VoxCeleb and LCFSH. [35] proposed an aging calibration method to compensate for the detrimental impact of aging on speaker verification performance. However, there is no large-scale dataset available for CA-SV studies. Currently, we found that celebrity audio-visual resource is inherently cross-age. Therefore, we mine cross-age test sets based on the VoxCeleb dataset. [36], [37] gathered age information from VoxCeleb, with [37] reporting 14,247 videos with age labels in VoxCeleb2, and [36] reporting 21,678 videos with age labels in VoxCeleb2 as well. However, considering the total number of videos in VoxCeleb2 development set (143,124 videos) and

VoxCeleb 1 (22,496 videos), [36], [37] only cover a very small portion. This limitation prevents for sufficient model training, especially when compared with the current published studies that were trained on Vox2dev and evaluated on VoxCeleb1. Therefore, we employ a facial age estimation method to label all videos in VoxCeleb1 & 2. Specifically, the paper constructs multiple cross-age test sets on VoxCeleb1 (Vox-CA), deliberately selecting positive trials with significant age gaps. The baseline system's performance experienced a noticeable drop from a 1.939% Equal Error Rate (EER) on the Vox-H [38] test set to 10.419% on the Vox-CA20 test set, as detailed in Section V-B1. Inspired by related works of face recognition [39], [40], [41], [42], we propose the Age Decoupling Adversarial Learning (ADAL) module to encourage speaker identity features to have smaller intra-class variations and be less correlated with the age information.

This paper builds on our previous work on cross-age speaker verification [43], while also introducing a novel short-term time-varying speaker verification task. Our previous work focused on the construction of the Vox-CA benchmark and the description of the learning age-invariant representation method, but due to space constraints, we were unable to provide comprehensive experimental details. In this paper, we provide more experimental details to fully present our findings of CA-SV and propose a new task with its corresponding solutions of STTV-SV. Our main contributions are as follows:

- A systematic discussion of time-varying speaker verification, including both short-term time-varying and cross-age speaker verification.
- Providing additional design details of the previously proposed cross-age speaker verification scenario.
- Releasing the SMIP-TV dataset, which focuses on the short-term time-varying speaker verification. Based on this dataset, we propose time-delay score and time-delay EER as auxiliary metrics to evaluate the ASV system over time.
- Introducing the incremental sequence-pair speaker verification task in the short-term time-varying scenario.
- Formulating the template updating process as a Markov Decision Process and using a deep reinforcement learning-based mechanism to determine the updating strategy.

The remaining paper is organized as follows. We discuss related works and introduce time-varying speaker verification tasks in Section II. Section III describes cross-age speaker verification task and proposed methods. Section IV presents the SMIP-TV dataset and our proposed deep reinforcement learning based template updating method. Sections V and VI provides the experimental setup and results analysis for CA-SV and STTV-SV, respectively. Finally, our conclusions are presented in Section VIII.

II. RELATED WORKS AND TASK INTRODUCTION

A. Related Works

In this part, we provide a brief overview of time-varying speaker verification, focusing on two categories: short-term and long-term time-varying speaker verification.

Short-term time-varying speaker verification: Early research has raised concerns about the impact of time-varying effects on speaker verification. [13] observed a decline in performance as the time gap between training and test data increased. [16] demonstrated that system performance dropped by nearly 50% when the time interval between enrollment and testing increased from 10 minutes to 2 months. Similarly, DeepSpeaker [19] reported a decrease in performance from 2.11% Equal Error Rate (EER) to 2.50% and 2.76% when the enrollment and testing intervals vary from 1 week to 1 mo and 3 months, respectively. In our summary report of Far-field Speaker Verification Challenge 2022 (FFSVC2022) [9], we also reported a significant performance decline with longer intervals between enrollment and testing audios.

Several datasets, such as CSLT-Chronos [15], CSLU [44], and TRSD [45], have been collected to investigate the effects of short-term time-varying variabilities. However, due to challenges in collecting time-varying data, most of these datasets have a limited number of speakers, such as 60 speakers in [15], 91 speakers in [44], and 55 speakers in [45]. Moreover, these datasets were recorded at specific time intervals rather than in continuous, real-life scenarios that encompass the complexities of varying conditions throughout the day. Additionally, while there are speech datasets available with speaker identity information, they are typically recorded only once a day, such as Automatic Speech Recognition (ASR) databases like LibriSpeech [46], AISHELL-2 [47], and TIMIT [32]. This limited temporal coverage may be one reason why the addition of ASR data has shown limited improvement on the speaker verification task [48], [49]. Therefore, there is a need for a dataset that covers a multiple randomly prompted time slots in each day within a continuous multi-month period from a large size cohort of speakers, capturing various states in real life from waking up to sleeping and aligning more closely with practical application scenarios.

To mitigate the negative effects of time-varying variations, red [15] proposed two modified acoustic features: pre-filtering frequency warping and post-filtering filter-bank outputs weighting, to alleviate the impact of time-varying factors. [50] proposed setting a prior decision threshold for speaker verification and provided examples of modifying the threshold during verification process to improve performance. However, this method requires continual operation for threshold tuning, which may not be practical in real-world applications. [24] discussed the impact of template aging in speaker verification and attempted to answer how often voice biometric templates should be updated. This viewpoint is similar to our proposed solution; however, due to data limitations, the author did not delve into further exploration and investigation.

Long-term time-varying speaker verification: [14] indicates that voice changes over time and samples with long-term gaps represent a challenge for speaker verification. To address this, the TCDSA database was introduced in [22], which contains long-term data from 18 speakers spanning a range of 30-60 years for each speaker. Based on the findings from TCDSA, authors of [17], [22], [23], [51] have concluded that using a decision threshold fixed at the time of enrollment results in a high

classification error rate after only a few years. They also observe that the issues of aging and quality variation are interconnected, with the effect of aging increasing over time and variations in quality becoming more likely. To address the associated variability from aging and quality, a verification decision boundary is proposed in score-aging-quality space by combining aging information with quality measures and the scores from the GMM-UBM system [51]. However, tuning the threshold lower may result in a decline in the miss rate, but the false alarm rate will rise, and vice versa. Although research on speaker verification in the context of aging is limited, age-invariant representation learning has been extensively studied in face recognition [39], [40], [41], [42] using model-wise approaches. Given that the aging process indeed increases intra-class variance, employing angle margin-based loss functions [6], [52] is also a reasonable method to handle it. However, age-relevant variabilities may not be specifically emphasized in loss-wise approaches.

B. Task Introduction

The traditional ASV task involves determining whether the claimed identity of an utterance matches a target identity. To evaluate the performance of a speaker verification system, a list of trial pairs is provided. Each trial pair consists of two single speech segments, and the scoring of each trial pair is independent. In the general task setting, as shown in Fig. 1(a), the trials are symmetric, meaning that the order of enrollment and test audio can be reversed without affecting the results. However, since this paper focuses on time-varying speaker verification, we take the chronological order into account when constructing the trial files. Therefore, we introduce two types of tasks: “Cross-age Single Pair-wise Speaker Verification” for long-term scenarios and “Incremental Sequence-pair Speaker Verification” for short-term scenarios of time-varying speaker verification.

1) *Cross-Age Single Pair-Wise Speaker Verification:* The traditional ASV scenario is considered a form of “Single pairwise speaker verification” where each trial consists of an enrollment audio and a test audio. In this paper, we consider the time factor, meaning that the enrollment audio is recorded earlier than the test audio. In FFSVC-22 [9], we intentionally varied the time intervals in test trials, where the first recording is used for enrollment and tested against the first, second, and third recordings. The results showed that as the recorded time interval increases, system performance decreases. Therefore, in this study, we aim to investigate the effects of larger time intervals and propose the “cross-age single pair-wise speaker verification task” as shown in Fig. 1(b). In this task, each trial is single pair-wise, where the positive pair of enrollment and test audio samples are selected in strict chronological order. For example, in our specific example (Fig. 1(b)), the test audio is captured at the present time, while the enrollment audio is obtained from a recording made ten years earlier.

2) *Incremental Sequence-Pair Speaker Verification:* The previous works on short-term time-varying scenarios [15], [16], [50] focused on the single pair-wise task. However, considering the variation over time of speech signals, we propose a novel and

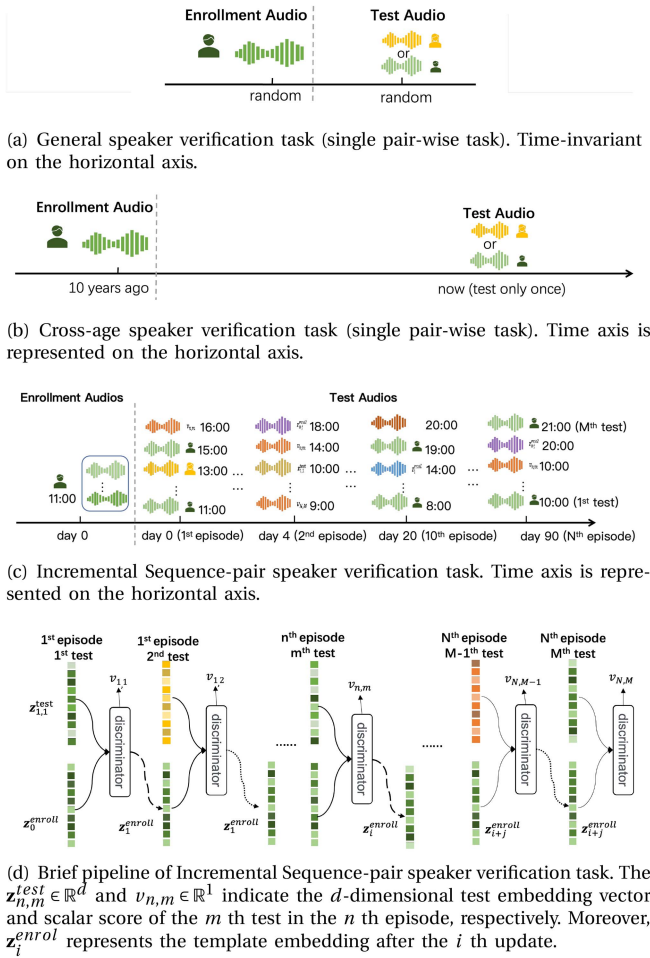


Fig. 1. Task introduction and schematic diagram of different ASV tasks. Different colored waveforms represent different speakers.

realistic task called “incremental sequence-pair speaker verification (ISSV)” to investigate short-term time-varying scenarios in speaker verification. Fig. 1(c) shows the form of the ISSV task. Compared to the single pairwise trial, the ISSV introduces two key differences. Firstly, each trial in ISSV consists of one enrollment template and multiple test audios. These test audios are sequentially presented to interact with the enrollment template, and their scores are evaluated in a chronological order. Secondly, the ISSV approach allows for performance improvement by incrementally updating the enrollment template during continuous interaction. Specifically, Fig. 1(d) illustrates a trial instance. Initially, the enrollment template z_0^{enroll} is generated as the average embedding extracted from the enrollment audios. This template embedding then interacts sequentially with the chronological testing embeddings, forming a trajectory. Within this trajectory, the template is scored against each test embedding. Subsequently, the discriminator module makes an identity decision based on the scores and updates the template embedding accordingly. The objective of this task is to evaluate the ASV systems’ ability to handle time-varying variabilities in real-world application scenarios.

III. LONG-TERM SPEAKER VERIFICATION

A. VoxCeleb Cross-Age Test Set

1) *Construction Details:* VoxCeleb1 [38] is a benchmark dataset for speaker verification, consisting of the original VoxCeleb1 original test set (Vox-O), the VoxCeleb1 extended test set (Vox-E), and the VoxCeleb1 hard test set (Vox-H). Vox-E is an evaluation protocol covering the entire dataset with 1,251 speakers. Vox-H is another evaluation protocol in which all negative pairs are from the same nationality and gender. We construct the Cross-Age test sets on VoxCeleb, named Vox-CA, which includes positive pairs with a large age gap and negative pairs of the same nationality and gender. The construction pipeline adopts the following steps:

- Gathering the face images from the meta-data of VoxCeleb1² and VoxCeleb2.³
- Estimating the age of each face image.
- Labeling the estimated age value for each audio utterance.
- Selecting positive pairs with a large age gap and negative pairs with speakers of the same nationality and gender.

For clarity, the key stages are described as follows:

Estimating and labeling age for audio: We use Dex [53],⁴ the winner in visual age estimation track of the ChaLearn LAP2015 challenge [54], to estimate the age value for each face image. Furthermore, as training set of Dex is derived from IMDB-WIKI and overlaps with speakers in VoxCeleb, Dex based age prediction is more accurate on VoxCeleb’s facial data. Since the audio of each utterance corresponds to multiple face images, the average age value of faces is used as the estimated age for this utterance. In addition, all the utterances of the same video segment should share the same age. Thus, the segment age, the average age among all the utterances belonging to the same video segment, is determined as the final age label. The estimated age distribution is shown in Fig. 2(a). In our calculation, the correlation between age labels in the Age-VOX-Celeb [36] dataset and our estimated age values is 0.83, with a mean absolute error (MAE) of 7.74. The 0.83 correlation coefficient result indicates that the predicted ages aligns well with the trend of actual ages. Although our estimated ages tend to be slightly older than the actual age values but the relative age gap and age range information are still captured. We create trials up to 20 years gap which could greatly tolerate the error in speaker age estimation.

Forming positive/negative pairs: Since the VoxCeleb2 dataset is typically used for training in most ASV system, the entire VoxCeleb1 dataset contributes to the construction of the cross-age test set with the following rules.

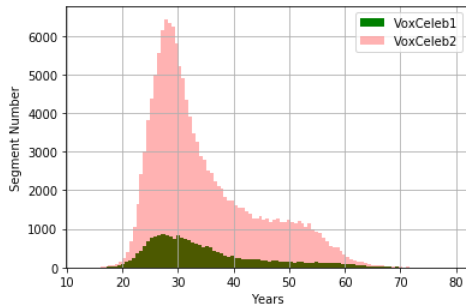
First, positive pairs must involve speakers from different age; i.e., the pair audios cannot be from the same video segment. We count the maximum age gap⁵ of each speaker and present the

²[Online]. Available: <https://www.robots.ox.ac.uk/~vgg/research/CMBiometrics/>

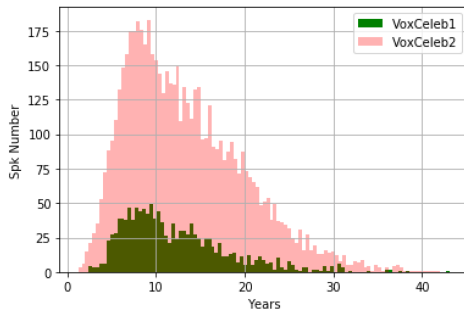
³[Online]. Available: <https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html>

⁴[Online]. Available: <https://data.vision.ee.ethz.ch/cv1/rrothe/imdb-wiki/>

⁵maximum age gap of one speaker indicates the difference between the largest estimated age value and the smallest estimated age value among all audio files of this speaker



(a) Distribution of estimated speaker ages among all segments in VoxCeleb 1&2



(b) Distribution of maximum age gap in terms of estimated ages among all speakers in VoxCeleb 1&2

Fig. 2. Statistics of speaker age and maximum age gap information in the VoxCeleb1&2 dataset based on the estimated age values.

distribution in Fig. 2(b). It is observed that the largest age gap for most speakers is between 0 and 20 years in VoxCeleb1, and only a few speakers have an age gap greater than 20 years. However, using too few speakers may affect the accuracy of the evaluation system, so the number of evaluation speakers must be taken into account when constructing the test set.

Second, following the Vox-H setting, we construct all negative pairs within the same nationality and gender. We maintain the same setting as Vox-H, where each nationality-gender combination has at least five individuals.

According to the rules mentioned above, we construct four Vox-CA sets according to different age-gap categories:

- Vox-CA5: The age gap of the positive pair is at least **5** years, and the candidate speakers must have more than 7 years of max age-gap data.
- Vox-CA10: The age gap of the positive pair is at least **10** years, and the candidate speakers must have more than 12 years of max age-gap data.
- Vox-CA15: The age gap of the positive pair is at least **15** years, and the candidate speakers must have more than 17 years of max age-gap data.
- Vox-CA20: The age gap of the positive pair is at least **20** years, and the candidate speakers must have more than 22 years of max age-gap data.

The Vox-CA test sets exhibit a progressive overlapping relationship, wherein Vox-CA5 may include speakers from Vox-CA10 and even Vox-CA20, thereby potentially encompassing

TABLE I
THE STATISTICS OF THE VOXCeleb1 TEST SET AND VOX-CA

Test set	Spk. Num.	Trials Num.	Age-gap	
			Positive	Negative
Vox-O	40	37611	2.68 ± 2.88	15.50 ± 12.46
Vox-E	1251	579818	3.14 ± 3.48	12.05 ± 9.81
Vox-H	1190	550894	3.14 ± 3.47	11.27 ± 9.42
Vox-CA5	971	370540	9.98 ± 3.94	12.36 ± 9.58
Vox-CA10	506	151384	15.29 ± 3.44	14.66 ± 9.93
Vox-CA15	215	54608	20.39 ± 3.38	16.63 ± 10.24
Vox-CA20	85	18888	25.28 ± 2.87	18.42 ± 10.58

Trials Num. and Spk. Num. Describe the number of trials and enrollment speakers, respectively. The column of positive and negative present the mean and standard deviation of age-gap values in corresponding pairs.

speakers with significant age gaps. However, it is essential to note that the construction process for Vox-CA5/10/15/20 is entirely independent and does not deliberately involve overlap. All the trials mentioned above have been released.⁶ The Vox-CA provides a challenging task that covering cross-age, same nationality and same gender cases. In addition, we also implement the single variable test set, including but not limited to: 1) test set within the cross-age; 2) test set within the same nation; 3) test set within the same gender; 4) test set within the intra-segment, to observe the effect of various factors on verification. The results are reported in Section V-B1.

2) *Comparison of Vox-E, Vox-H and Vox-CA*: In this part, we compare the difference of Vox-E, Vox-H and Vox-CA from various aspects.

Positive pair within the cross-age and intra-segment: The average age gap of the VoxCeleb1 test set is approximately 3 years, as shown in Table I. Considering the error in age estimation, most positive trails are from the similar period. The positive pairs in the VoxCeleb1 test set are chosen randomly from the same person without considering the age gap. However, Vox-CA intentionally selects pair audio from larger age-gap segments. The other extreme is when the pair of audios are chosen from the same video segment, resulting in a higher successful verification rate [43].

Negative pair within the same nationality and gender: Both Vox-H and Vox-CA take nationality and gender into account when constructing negative pairs. In contrast, Vox-E randomly selects pairs from the entire dataset. Thus, the Vox-H and Vox-CA sets are more challenging.

Overall, the Vox-CA test sets provide a more challenging evaluation for speaker verification systems by introducing larger age gaps, while also considering nationality and gender constraints for negative pairs. This enables a comprehensive assessment of the system's performance under cross-age, same nationality, and same gender scenarios.

B. Learning Age-Invariant Speaker Embedding

1) *Toy Experiment*: The results of the cross-age scenario (only-CA) in Table V present that as the time gap increases, the performance decrease. This observation leads to the research

⁶[Online]. Available: https://github.com/qinxiaoyi/Cross-Age_Speaker_Verification

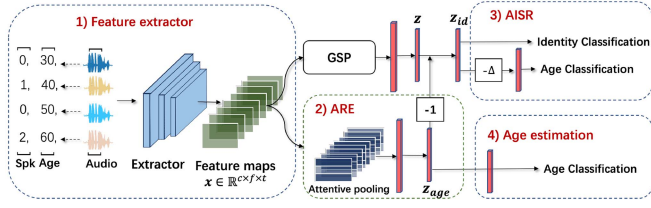


Fig. 3. An overview of the proposed ADAL structure. The AISR denotes the age-invariant speaker representation. GSP is denoted as global statistic pooling layer.

question of whether speaker embedding contains age information. To investigate this, a toy experiment is conducted by using a pre-trained ASV system to extract speaker embeddings and predict the speaker age. The age classifier is employed to classify the age into 7 groups: 0–20, 21–30, 31–40, 41–50, 51–60, 61–70, and 70–100. The speaker embeddings are fed into a linear layer for age class prediction, achieving an accuracy of 82.01%. This high accuracy indicates that speaker embeddings indeed contain age information. Therefore, the goal is to learn an age-invariant speaker embedding to mitigate the negative effect of age.

2) *Decoupling Age-Related Component*: The assumption is made that the speaker embedding consists of identity and age information driven by their respective tasks. To decouple the age information from the identity features, a linear model is designed. Specifically, the feature embedding $\mathbf{z} \in \mathbb{R}^d$, a d -dimensional vector extracted from an input audio, is assumed to be the sum of the identity component \mathbf{z}_{id} and the age component \mathbf{z}_{age} [40]:

$$\mathbf{z} = \mathbf{z}_{id} + \mathbf{z}_{age} \quad (1)$$

An Age-Related Extractor (ARE) module is introduced to extract age-related information from the high-level feature maps $\mathbf{x} \in \mathbb{R}^{C \times F \times T}$, where C , F , and T indicate the dimensions of the channel, frequency and temporal domains, respectively. The ARE module, utilizing the attention mechanism, includes a pooling layer (*pool*), a fully connected layer (*fc*), and an attention module denoted by σ . The age-related embedding with d -dimensions is obtained by applying pooling, linear transformation, and attention to the output of the attention module, as expressed by the following equations:

$$\mathbf{z}_{age} = ARE(\mathbf{x}) = fc(pool(\mathbf{x} \odot \sigma(\mathbf{x}))) \quad (2)$$

Here, Attentive Statistical Pooling (ASP) [55] is utilized to perform the $pool(\mathbf{x} \odot \sigma(\mathbf{x}))$ operation. The purpose of pooling is to transform variable-length speech features into fixed-length vector representations along the temporal direction.

Then, the age-related component \mathbf{z}_{age} is subtracted from \mathbf{z} , effectively reducing the age-related information for supervision by an age classifier.

3) *Multi-Task Learning*: Fig. 3 provides a detailed overview of the proposed network structure. We adopt multi-task learning, which involves three supervised tasks: identity classification, age classification, and age adversarial learning.

Identity classification: We adopt the Identity Classifier layer (*IC*) to guide the \mathbf{z}_{id} to represent the identity information.

To account for speaker aging and the resulted large intra-class variance in CA-SV, ArcFace is employed as the identity loss function to reduce intra-class distance.

Age classification: To decouple the age information from the speaker embedding, an age classifier A is employed to supervise the learning of age-related embeddings. In general, the combination of age-classification and regression loss is adopted as loss function for age estimation. However, since age values are estimated by faces and not directly obtained from ground truth, the estimated age labels contain noise. Therefore, an age group classifier is used, where the age groups correspond to the ones used in Section III-B1.

Age adversarial learning: To further reduce the age information contained in the identity embedding \mathbf{z}_{id} , an additional age classifier with gradient reversal layer (GRL) [56] is applied upon the \mathbf{z}_{id} .

The proposed method is named as Age Decoupling Adversarial Learning (ADAL). The final loss function for the method is formulated as follows:

$$\mathcal{L}_{id}(\mathbf{z}_{id}) = l_{ce}(IC(\mathbf{z}_{id}), y_{id}) \quad (3)$$

$$\mathcal{L}_{age}(\mathbf{z}_{age}) = l_{ce}(A(\mathbf{z}_{age}), y_{age}) \quad (4)$$

$$\mathcal{L} = \mathcal{L}_{id}(\mathbf{z}_{id}) + \lambda_{age} \mathcal{L}_{age}(\mathbf{z}_{age}) + \lambda_{grl} \mathcal{L}_{age}(GRL(\mathbf{z}_{id})) \quad (5)$$

where $y_{id} \in \{0, 1, \dots, N\}$ and $y_{age} \in \{0, 1, \dots, 6\}$ are the output labels of identity and age estimation tasks, respectively. l_{ce} denotes the cross-entropy loss, and λ_{age} and λ_{grl} are scalars used to balance different loss terms.

IV. SHORT-TERM TIME-VARYING SPEAKER VERIFICATION

A. SMIIIP-TimeVarying Dataset

The SMIIIP-TimeVarying Dataset (SMIIIP-TV), is a speaker verification dataset designed for research purposes that focuses on short-term time-varying of speaker verification. The recordings language is mandarin. The dataset contains recordings from 373 speakers who provided utterances over 90 consecutive days, in which each speaker needs to record multiple utterances at varying time slots in each day. To ensure that recording time spans the full day without location limitations, we developed an Android application, which randomly assigns recording tasks in five different time slots: 6:00-8:00, 9:00-11:00, 12:00-14:00, 17:00-19:00, and 20:00-22:00, as shown in Fig. 4(a). In each time slot, speakers provide three utterances, including both text-dependent and text-independent speech samples. Table II shows the contents of the recording. Additional meta information such as speaker region (in total 27 provinces, China), age, and cell-phone type were collected. Additionally, speakers were asked to report details on their physical state (in total 7 types, including normal, sleepy, eating, sore throat, exercise, cold/fever, others), recording environment (in total 16 scenes) and the degree of noise (in total 4 levels, including quiet, normal, noisy, extremely noisy), all were manually reviewed. The dataset statistics are presented in Fig. 4. The majority of speakers in the dataset are college students and their families from Shanxi Province, China, and the gender distribution is balanced (171 males:202

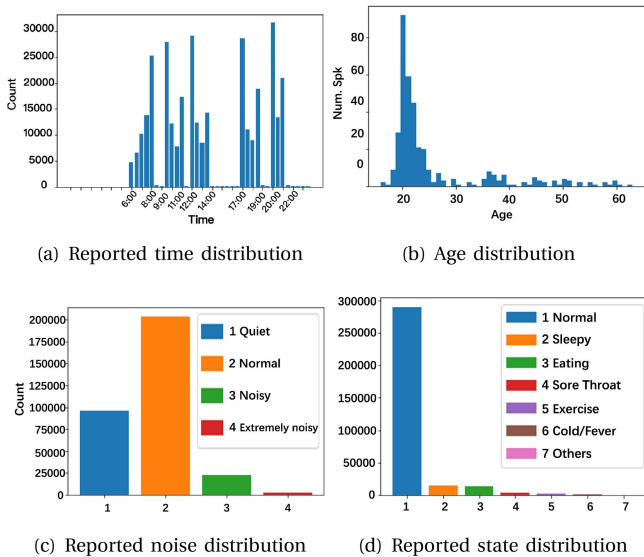


Fig. 4. Facts of the SMIP-TV dataset in terms of data collection time, speaker age, background noise and user states.

TABLE II
THE CONTENT STATISTICS OF THE SMIP-TV DATASET

Content	Num.	Average Duration (w/o VAD)
'ni hao, mi ya'	27138	1.86s
'xiao le, xiao le'	26923	1.88s
'xiao ai tong xue'	27227	1.92s
'tian mao jing ling'	27076	1.91s
'tong li tong li'	26972	1.88s
free text	189713	4.45s
total	325049	3.38s

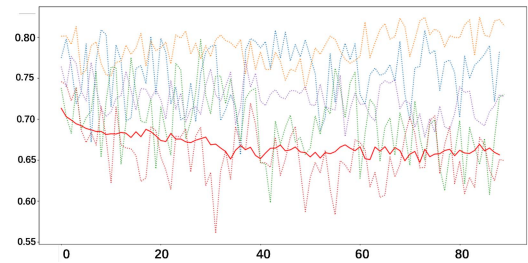
Content and Num. describe text content and its corresponding utterance number, respectively.

females). Most recordings were made indoors, with majority of the noise and physical conditions being normal. Speakers were also encouraged to report various scenes with different physical conditions. Due to the challenge of continuously recording for 90 days, some speakers were unable to provide recordings for the entire duration. Finally, 133 speakers recorded for the entire 90-day period, and we selected 58 of them as the SMIP-TV test set, and the remaining speaker data (315 speakers) is adopted as the training set. The entire dataset is available⁷ for publicly releasing.

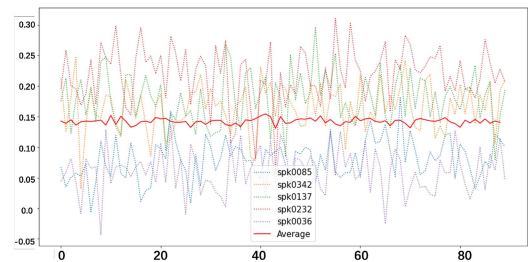
B. Analysis of Time Varying

The aim of this study is to examine speaker variability over several days, which is referred to as Short-term Time-varying Speaker Verification. First, the enrollment template embedding of each speaker is the average of embeddings from corresponding audio samples of the first day, which is then tested against audio samples from the second, third, and subsequent days up to the N_{th} day. By analyzing the changes in EER and scores curve

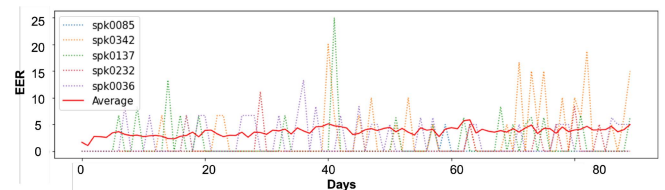
⁷https://github.com/qinxiaoyi/TimeVarying_ASV



(a) Positive trial score curve of each day from 5 random speakers.



(b) Negative trial score curve of each day from 5 random speakers.



(c) EER curve of each day from 5 random speakers

Fig. 5. Results in each day scenario. “Average” indicates the average value of all test speakers in each day. Three subfigures share one legend.

for each individual on a daily basis, insights can be gained into the temporal dynamics of speaker verification. However, due to the requirement of a large number of positive and negative trial scores for calculating EER, and on average only 14.7 speech samples are available for each speaker per day in our test set, there are significant fluctuations for both positive trial scores and EER as shown in Fig. 5.

Therefore, we propose two metrics for analyzing time varying: the Time-delay Score (TDS) curve and the Time-delay EER (TD-EER) curve. The TDS is the average scores of current day and previous days, and the TD-EER computes the EER based on the positive and negative TDS. These metrics are combined with the overall EER to evaluate the impact of time varying. The specific definitions are as follows:

First, we make a definition that in a given set $A \in \mathbb{R}$, the average of set A is denoted as

$$\bar{A} = \frac{1}{|A|} \sum_{a \in A} a$$

Then, set $S_{i,j}$ is a subset of trial scores from the j_{th} speaker on the i_{th} day (total N speaker with M days). The set of all scores

up to k day is denoted as

$$TD_k := U_{i \leq k, j \leq N} S_{i,j}$$

The set of positive TDS (TDS^p) of the j th speaker can be formulated as

$$TDS_j^p(k) = \overline{TD_{k,j}^p}$$

The subset of $TD_{k,j}^p := U_{i \leq k} S_{i,j}^p$, where trial ground truths are positive. The superscript p and n indicates the positive and negative trial scores, respectively.

$TDS^p(k)$ and $TDS^n(k)$ denote the collections of all positive trial scores and all negative trial scores up to the k th day, respectively. TDS^p and TDS^n curves against the time information are presented in Fig. 6(a) and (b).

Therefore, the curve function of TD-EER (shown in Fig. 6(c)) can be formulated as :

$$TD - EER(k) = EER(TDS^p(k), TDS^n(k)) \quad (6)$$

Additionally, we plot the sliding window EER (SW-EER) curve in Fig. 6(d), which computes the EER based on the positive and negative trial scores over a window period that slide along the time axis to observe the short-time system performance. We use a window length of 10 days and a hop length of 1 d in this study.

The results depicted in Fig. 6 indicate a gradual decline trend in the TDS^p curve and a slowly increasing trend in the TD-EER curve, indicating a degrading pattern in term of system performance over time. These trends are in line with the findings in [15]. Moreover, the negative trial score (TDS^n) curve exhibits fluctuations within a limited range, indicating that time varying has little impact on negative samples.

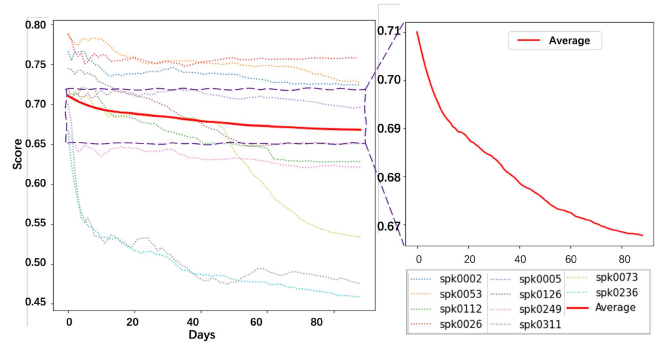
Based on the aforementioned observations and analyses, we can conclude that there is a slow effect of time varying on positive trial scores, while negative trials are relatively unaffected. As the positive trial scores decline over time, the system performance will also degrade. This raises the question of how to maintain or even enhance the performance of the speaker verification systems over time when we continuously use it, e.g. mobile phone login.

C. Instance-Wise Template Updating

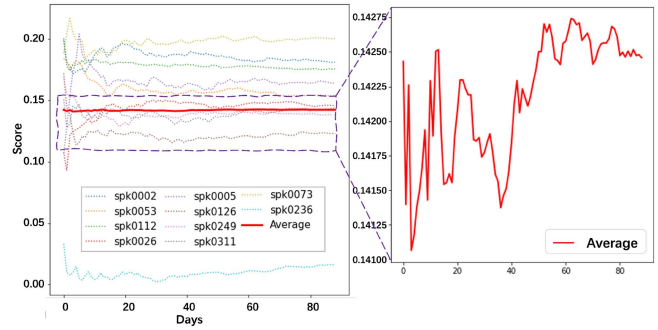
The analysis of short-term time varying (depicted in Fig. 6) indicates a noticeable decline in performance within a three-month period. To address this issue, we propose an instance-wise template updating approach that updates the template after each validation, reducing the need for long-term storage and personalized model creation. This method is particularly suitable for situations where privacy concern associated with storing audio samples over extended periods is an issue or developing customized models for individual users is challenging.

1) *Fixed-Weight Template Updating Method*: Firstly, we introduce a fixed-weight template updating method (FixW-TU) that updates the enrollment template sequentially when the test audio progresses, as described in Section II-B2. The updating method is introduced in [20]:

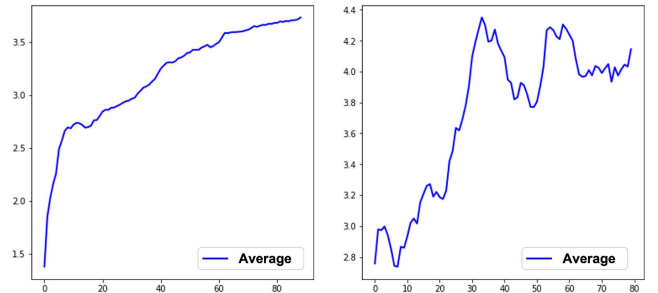
$$\mathbf{z}_{t+1}^{enroll} = (1 - \alpha) * \mathbf{z}_t^{enroll} + \alpha * \mathbf{z}_t^{test} \quad (7)$$



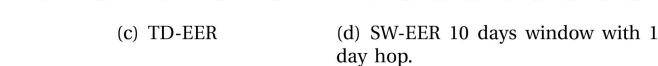
(a) TDS^p curve of 10 random speakers.



(b) TDS^n curve of 10 random speakers.



(c) TD-EER



(d) SW-EER 10 days window with 1 day hop.

Fig. 6. Results of TDS curves, TD-EER, and SW-EER on the SMIP-TV test set. The exemplar speakers are randomly sampled from the test set, with the “average” curve representing the average results across all 58 individuals in the test set.

where α represents a predefined updating weight, and \mathbf{z}^{enroll} and \mathbf{z}^{test} are d -dimensional enrollment template and test embedding, respectively. The template updating is triggered when the similarity score between enrollment and test embedding exceeds the predetermined updating threshold. The threshold is determined through multiple manual evaluations. Algorithm 1 provides details of the FixW-TU method.

2) *Reinforcement Learning Based Template Updating Method*: The FixW-TU approach is limited by the need of manual calibration to identify appropriate values for the updating threshold β and the update weight α . Although we can achieve relatively good results in a specific scenario through manually tuning the parameters, the optimal parameters might

Algorithm 1: FixW-TU With Thresholding.

```

//Initialization
1. Enrollment template: Random sample  $K$  embeddings
 $\mathbf{z}_0^{enroll} \in \mathbb{R}^{K \times d}$  from a speaker and calculating the average
to obtain the template  $\mathbf{z}_0^{enroll}$ .
2. Test set: Randomly sampling negative samples and
sequentially sampling positive samples, forming a  $T$ -length
sequence  $\mathbf{z}^{test} = \{\mathbf{z}_0^{test-p}, \mathbf{z}_1^{test-n}, \dots, \mathbf{z}_{T-1}^{test-p}\}$ . The ratio of
positive and negative samples is 1:1.
3. Fixed weight:  $\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5\}$ 
// Implementation
4. Determine updating threshold :  $\beta=0.51$  (empirical value)
5. Terminated=0
6. while not Terminated do
    $v_t = \text{cosine}(\mathbf{z}_t^{enroll}, \mathbf{z}_t^{test})$ 
   if  $v_t > \beta$  then
      $\mathbf{z}_{t+1}^{enroll} = (1 - \alpha) * \mathbf{z}_t^{enroll} + \alpha * \mathbf{z}_t^{test}$ 
   else
      $\mathbf{z}_{t+1}^{enroll} = \mathbf{z}_t^{enroll}$ 
   Terminated=1 if  $t+1==T$ 

```

vary in different setups and using cases. Therefore, we are motivated to explore adaptive updating methods where the updating decision and the updating weight can be automatically calculated in a customized manner. We found that template updating is a sequential decision-making problem with the goal of maximizing long-term benefits. Therefore, we propose a Deep Reinforcement Learning-based template updating (DRL-TU) method that formulates the problem of finding suitable thresholds and weights as a Markov Decision Process (MDP), described by $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$ as the states, actions, transitions, and rewards. Our approach adopts the Proximal Policy Optimization (PPO) [57] strategy to optimize an agent and aims to achieve durable verification benefits for the overall system. In this paper, we propose two DRL-based template updating methods, namely DRL Adaptive Weighted Template Updating with thresholding (DRL-TU-AdW) and DRL Multi-Head Template Updating (DRL-TU-MH). Next, we will present the interaction environment and the algorithm in details.

3) *Interaction Environment*: Our proposed DRL-TU methods operate within an interaction environment consisting of States, Actions, Transitions, and Rewards, as illustrated in Fig. 7.

States: The input state $s_t \in \mathcal{S}$ for the agent is composed of $\{\mathbf{z}_t^{enroll}, \mathbf{z}_t^{test}\}$, where \mathbf{z}_t^{enroll} and \mathbf{z}_t^{test} is d -dimensional enrollment and test speaker embeddings, respectively.

Actions and transitions: For the DRL-TU-MH method, in every state, the agent can take two actions: the binary updating decisions a_t^{act} and the corresponding updating weight a_t^{weight} . The weight a_t^{weight} is a scalar that provides the updating weight for the test embedding. The updating decision a_t^{act} belongs to the set $\{0, 1\}$, where $a_t^{act} == 1$ indicates that the template needs to be updated, while $a_t^{act} == 0$ indicates that it does not need to be updated. On the other hand, the DRL-TU-AdW method only predicts the updating weight a_t^{weight} , and the updating decision is made by setting a threshold. Before the termination of each trajectory, the agent transits to the next state based on the transition distribution $\mathcal{T}(s_{t+1}|s_t, a_t)$.

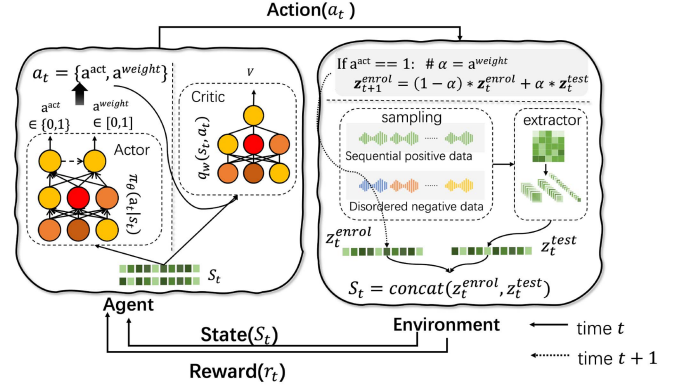


Fig. 7. The brief training structure of our DRL based template updating method. This figure presents the pipeline of the DRL-TU-MH method. The left part signifies the forward direction in one training iteration. The enrollment embedding and randomly selected test audio embedding are concatenated to form a state. This state undergoes state transition through the agent, resulting in an action. Following the action, a new enrollment embedding is obtained, constituting the new state in next stage.

TABLE III
REWARD MATRIX IN ALGORITHM 2

	$a_t^{act} == 1$	$a_t^{act} == 0$
Label==1	$r_t^{pair} + r_t^{cen} + 0.5$	$r_t^{pair} + r_t^{cen}$
Label==0	$-r_t^{pair} - 1$	0.5

Rewards The reward function of DRL-TU-MH is consisted of three parts: the accuracy of the updating decisions r_t^{dec} , cosine similarity between the updated enrollment embedding and the next-stat test embedding r_t^{pair} , and the cosine similarity between the updated enrollment embedding and the speaker center embedding r_t^{cen} . The reward r_t for each action is provided in Table III.

The r_t^{pair} and r_t^{cen} are formulated as follows:

$$r_t^{pair} = \cos(\mathbf{z}_{t+1}^{enroll}, \mathbf{z}_{t+1}^{test}) \quad (8)$$

$$r_t^{cen} = \cos(\mathbf{z}_{t+1}^{enroll}, \mathbf{z}^{cen}) \quad (9)$$

where $\cos(\cdot)$ indicates the cosine similarity between two input embeddings, and \mathbf{z}^{cen} represents the average embedding of samples collected over a 90-day period for the enrollment speaker. The $\mathbf{z}_{t+1}^{enroll}$ represents that the embedding has been updated using (7). Since the updating decision of DRL-TU-AdW is made by setting a threshold, reward function only adopts the r_{pair} and r_{cen} to guide the agent. Therefore, the T -length trajectory in each episode can be indicated as the iteration set of $\{(s_t, a_t, s_{t+1}, r_t)\}$. The interaction details of training are summarized in Algorithm 2.

4) *PPO Based Template Updating*: The PPO algorithm is used to develop an optimal updating policy for the agent to create an enrollment template that can adapt automatically for better short- and long-term benefits. It is well known that supervised learning may not be suitable for the sequential stochastic and decision-making problem involved in template updating. Therefore, we employ DRL to address this issue. We have explored Deep Deterministic Policy Gradient (DDPG) [58] and

Algorithm 2: DRL-TU Interaction Environment Under the Training Stage.

```

// Initialization
1. Enrollment template: Random sample  $K$  embeddings
 $\mathbf{z}_0^{enroll} \in \mathbb{R}^{K \times d}$  from a speaker and calculating the average
to obtain the template  $\mathbf{z}_0^{enroll}$ .
2. Test set: Randomly sampling negative samples and
sequentially sampling positive samples, and forming a
 $T$ -length sequential test set
 $\mathbf{Z}^{test} = \{\mathbf{z}_0^{test-p}, \mathbf{z}_1^{test-n}, \dots, \mathbf{z}_{T-1}^{test-p}\}$  with its labels
 $\{l_0, l_1, \dots, l_{T-1}\}$ .
// Implementation of DRL-TU-AdW
3a. Determine threshold of updating:  $\beta=0.51$ 
4a. Terminated=0
5a. while not Terminated do
     $v_t = \text{cosine}(\mathbf{z}_t^{enroll}, \mathbf{z}_t^{test})$ 
    if  $v_t > \beta$  then
         $a_t = \pi_\theta(\mathbf{z}_t^{enroll}, \mathbf{z}_t^{test})$   $\alpha = \text{sample}(a_t)$ , where
         $a_t \sim \mathcal{N}(\mu, \sigma^2)$   $\mathbf{z}_{t+1}^{enroll} = (1 - \alpha) * \mathbf{z}_t^{enroll} + \alpha * \mathbf{z}_t^{test}$ 
    else
         $\mathbf{z}_{t+1}^{enroll} = \mathbf{z}_t^{enroll}$ 
     $r_t = \text{RewardMatrix}(s_t, a_t, l_t)$ 
    Terminated=1 if  $t+1 == T$ 
// Implementation of DRL-TU-MH
3b. Terminated=0
4b. while not Terminated do
     $a_t^{act}, a_t^{weight} = \pi_\theta(\mathbf{z}_t^{enroll}, \mathbf{z}_t^{test})$ 
     $action = \text{sample}(a_t^{act})$ 
     $\alpha = \text{sample}(a_t^{weight})$ , where  $a_t \sim \mathcal{N}(\mu, \sigma^2)$ 
    if  $action == 1$  then
         $\mathbf{z}_{t+1}^{enroll} = (1 - \alpha) * \mathbf{z}_t^{enroll} + \alpha * \mathbf{z}_t^{test}$ 
    else
         $\mathbf{z}_{t+1}^{enroll} = \mathbf{z}_t^{enroll}$ 
     $r_t = \text{RewardMatrix}(s_t, a_t, l_t)$ 
    Terminated=1 if  $t+1 == T$ 

```

Advantage Actor Critic (A2C) [59] to predict the updating weight and make updating decision, respectively. But the training process is unstable and the results are even worse than the FixW-TU baseline. Hence, we adopt the PPO strategy to train the policy-based agent in learning continuous and discrete action spaces. Here, we provide the details of two DRL-based methods: DRL-TU-AdW and DRL-TU-MH.

DRL-based Adaptive Weighted Template Updating with Thresholding (DRL-TU-AdW): This method utilizes DRL to predict the updating weight for each verification objective function. The weight is determined using a clipped version of PPO [57]:

$$\mathcal{L}^{policy}(\theta) = \min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t) \quad (10)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$. The policy π_θ is implemented using a neural network (NN)-based agent, and \hat{A}_t denotes the estimator of the advantage function at time step t . As PPO is an actor-critic algorithm which combines elements of both value-based methods (critic) and policy-based methods (actor) to improve the efficiency and stability of learning [57], [60], the final objective function of DRL-TU-AdW is a combination of policy and value

network, defined as follows:

$$\mathcal{L}^{PPO}(\theta) = \mathcal{L}^{policy}(\theta) + c_1 \mathcal{L}^{value}(w) \quad (11)$$

where $\mathcal{L}^{value}(w)$ represents the mean squared error (MSE) loss of the state-value function, and c_1 is a coefficient.

DRL based Multi-head Template Updating (DRL-TU-MH) improves upon the threshold-based DRL-TU-AdW method, which only considers values within boundary limits, thereby limiting the explorability of intra-class space and causing template trends to move towards the initial embeddings center. To address this issue, we adopt the Hybrid Action Space PPO [61], which can handle both continuous and discrete action spaces $a_t \in \{a_t^{act}, a_t^{weight}\}$. Specifically, we use a multi-head agent where the first action head outputs a categorical decision that determines whether to update the template or not. This output then feeds into subsequent action heads to learn the updating weight with a normal distribution. The objective function includes the entropy loss for both the categorical and normal distributions:⁸

$$\begin{aligned} \mathcal{L}^{entropy}(\theta) = & H_{discrete}(\pi_\theta^{dis}(s_t)) \\ & + H_{continue}(\pi_\theta^{con}(s_t, a_t^{act})) \end{aligned} \quad (12)$$

where $H_{discrete}(\cdot)$ and $H_{continue}(\cdot)$ are the Shannon entropy and differential entropy, respectively. $\pi_\theta^{dis}(s_t)$ indicates the categorical output of first action head, while $\pi_\theta^{con}(s_t, a_t^{act})$ is normal output of the section action head. The final objective function is:

$$\mathcal{L}^{PPO}(\theta) = \mathcal{L}^{policy}(\theta) + c_2 \mathcal{L}^{value}(w) - c_3 \mathcal{L}^{entropy}(\theta) \quad (13)$$

In contrast to DRL-TU-AdW, DRL-TU-MH adopts a two-stage training approach: a) supervised pre-training and b) reinforcement learning based fine-tuning. First, we use supervised pre-training to set the initial agent parameters. The input feature is the concatenation of two speaker embeddings (extracted by pre-trained model). The first head is a binary classifier that determine whether two embeddings are from the same person. The second head is a regression task and we setting the ground truth as $\alpha = 0.15$, which is the empirical value from the FixW-TU method. In the fine-tuning stage, we optimize the pre-trained model parameters using the PPO algorithm.

The design of the reward function and the weights of the objective function in DRL are determined through multiple trials on the development set of our dataset.

V. EXPERIMENTAL RESULTS OF CA-ASV

A. Implementation Details

1) *Network*: For the baseline system, termed ResNet34-ArcFace, we adopt the ResNet34 [62] as the backbone. The widths (channels number) of the residual blocks are {32, 64, 128, 256}. The global statistic pooling (GSP) layer, which computes the mean and standard deviation of the output feature maps, can project the variable length input to the fixed-length vector. The output of a fully connected layer with 128 dim followed after the pooling layer is adopted as the speaker embedding

⁸Implementation of multi_action_head_PPO github repository

TABLE IV
THE PERFORMANCE OF DIFFERENT SPEAKER VERIFICATION SYSTEMS IN TERMS OF EER

Model	Vox-E	Vox-H	Cross-age				Cross-age & Same nationality & Same gender			
			Only-CA5	Only-CA10	Only-CA15	Only-CA20	Vox-CA5	Vox-CA10	Vox-CA15	Vox-CA20
ResNet34-Softmax	2.798%	4.806%	4.310%	6.004%	8.019%	9.308%	7.366%	9.215%	12.405%	14.888%
ResNet34-ArcFace	1.094%	1.939%	1.953%	3.437%	5.927%	8.185%	3.407%	4.974%	8.028%	10.419%
+ GRL	1.122%	1.934%	2.021%	3.579%	6.036%	8.566%	3.405%	4.949%	8.017%	10.610%
+ Age Residual	1.121%	1.960%	2.040%	3.536%	5.871%	7.864%	3.499%	5.078%	8.039%	10.229%
+ ARE (ours)	1.108%	1.951%	1.980%	3.345%	5.719%	7.803%	3.431%	4.814%	7.786%	9.911%
+ ADAL (ours)	1.121%	1.974%	1.991%	3.330%	5.540%	7.442%	3.441%	4.822%	7.515%	9.519%

The model with GRL describes the simplest adversarial learning that uses GRL upon the z vector to perform the age classification task, which makes the speaker embedding less correlated to age. In the age residual method, z_{id} is the residual part between z and z_{age} , the z_{age} is extracted from z and supervised by age classification. The model equipped with ARE indicates the z_{age} is also supervised by age classification without age adversarial learning.

layer. The ArcFace-based classifier [6] ($s=64, m=0.2$), which increase intra-speaker distances while ensuring inter-speaker compactness, is used to the identity classification. In addition, we also provide the Softmax classifier as a comparison. For the ADAL method, the ASP system is adopted as the ARE module to extract the z_{age} vector. For the age classification, we stack FC-ReLU-FC structure upon z_{age} and z_{id} to predict the age group value.

2) *Data Processing*: The acoustic features are 80-dimensional log Mel-filterbank energies with a frame length of 25 ms and hop size of 10 ms. We adopt the on-the-fly data augmentation [63] to diversify training samples. Four types of augmentation methods were adopted: 1) adding noise using MUSAN [64] dataset; 2) adding convolutional reverberation using RIR Noise [65] datasets; 3) changing amplification, and 4) changing audio speed (pitch remains untouched).

3) *Training Details*: The SGD optimizer is employed to update the model parameters. We adopt the multi-step learning rate (LR) scheduler with 0.1 initial LR; the decay step and factor are 10 and 0.1, respectively. We adopt the linear warmup from 0.0 to 0.1 LR in the first two epochs to prevent the training instability and speed up model convergence. Training stopped after LR dropped to $1e-5$. In order to ensure that the model remains primarily focused on the task of speaker identity without compromising performance, the hyper-parameters in loss are set as following: $\lambda_{age} = 0.1$ and $\lambda_{grl} = 0.1$. We experimented with different coefficients, namely 0.5, 0.3, and 0.1, and observed that as the coefficient increased, the model's performance deteriorated on the general test sets Vox-E and Vox-H, while showing limited improvement on Vox-CA.

4) *Evaluation Measures*: Cosine similarity is used for trial scoring. Verification performances are measured by EER and the minimum normalized detection cost function (mDCF) with $P_{target} = 10^{-2}$ and $C_{FA} = C_{Miss} = 1$.

B. Experimental Results and Analysis

1) *Experimental Results of the Baseline Method*: In this part, we adopt ResNet34-GSP-ArcFace as our baseline system. We compare the baseline performance on Vox-O, Vox-E, Vox-H and our proposed Vox-CA. Table V reports the corresponding results which confirms the difficulty of the Vox-CA test set.

First, by observing the performance on our-E and our-H (our implemented following the VoxCeleb rules), the results are similar to Vox-E and Vox-H results, that demonstrate the correctness

TABLE V
RESULTS ON DIFFERENT TEST SETS BASED ON THE RESNET-GSP-ARCFACE MODEL

Test set	Variable	EER[%]	mDCF _{0.01}
Vox official			
Vox-O	random	0.962%	0.100
Vox-E	random	1.094%	0.122
Vox-H	nation&gender	1.939%	0.200
our proposed			
our-E	random	1.202%	0.123
our-H	nation & gender	2.044%	0.192
only-N	nation	1.568%	0.164
only-G	gender	1.534%	0.146
only-I	intra-segment	0.227%	0.015
only-CA5	age	1.953%	0.177
only-CA10	age	3.437%	0.272
only-CA15	age	5.927%	0.352
only-CA20	age	8.185%	0.464
Vox-CA5	age & nation & gender	3.407%	0.300
Vox-CA10	age & nation & gender	4.974%	0.370
Vox-CA15	age & nation & gender	8.028%	0.481
Vox-CA20	age & nation & gender	10.419%	0.646

"Only" indicates that a trial is created by only considering a single variable.

of our dataset construction. Then, by controlling a single variable, the negative effect of cross-age (only-CA) is larger than the same nationality (only-N) and gender (only-G) matching. When we combine these variabilities, the performances of Vox-CA drops dramatically with the age gap increasing. The Vox-CA not only provides a new hard scenario but also proposes a new benchmark for cross-age scenarios. In addition, the result of the intra-segment case is considerably lower than other test sets. The validation of intra-segment pairs is too easy, which can lead to misjudgment of the actual performance of the system.

2) *Experimental Results of AISR*: Table IV presents the performance of our proposed methods and related methods on different test sets. First, we compare different metric learning methods, namely Softmax and ArcFace. We can find that the ArcFace outperforms its counterpart, especially in cross-age scenarios. Besides, by comparing the results on cross-age test sets, we can observe that the verification performance degrades significantly with age-gap increasing. Using the ArcFace based system, we mount the GRL or Age Residual module as comparison. The model with GRL is an adversarial learning method which is the combination of identity classification and age adversarial learning in Section. The model with Age Residual module is the combination of identity classification and age

classification, but the \mathbf{z}_{age} is extracted from the \mathbf{z} . However, these methods have little improvement on cross-age scenarios. We think the limitation of both methods is that operations are performed on the embedding level. Since the embedding is a compact representation vector generated by the encoder layer, resulting in limited operational margins, thus the improvement is moderate. In contrast with these methods, the \mathbf{z}_{age} of ADAL are extracted from high-level feature maps, and age information is further reduced by the age adversarial learning classifier. In contrast to the baseline system, the ADAL achieves 10% relative improvement on the Vox-CA20 test set. Furthermore, the results show that the performance improves with larger age gaps. Finally, we utilized the embeddings learned by ADAL for age class prediction again, observing a decrease in the accuracy of age identification from 82.01% (as reported in Section III-B1) to 72.44%. This suggests that the embeddings learned by ADAL have diminished age-related information.

VI. EXPERIMENTAL RESULTS OF STTV-ASV

A. Implementation Details

1) *Network*: To perform template updating on the embedding level, we use the same pre-trained baseline model (ResNet34-ArcFace) as in CA-ASV and fine-tuned it with the SMIP-TV training set. Specially, we divide the SMIP-TV set into training and test sets. We randomly select 58 speakers who completed the entire recording as the test set to construct the sequential trials, and the data of the remaining speakers as the training set. The data processing is the same as CA-SV. The training set is divided into 5 folds for cross-validation to tune the decision threshold, determine hyper-parameters, and assess the performance of the DRL-TU system. The final system for speaker embedding extraction is based on fine-tuning a pre-trained ResNet34-ArcFace model using the Vox2dev and SMIP-TV training sets.

For the DRL-TU-MH method, we employed a two-layer fully connected structure for the agent model, which is pre-trained with supervision on the SMIP-TV training dataset. During the fine-tuning stage of DRL, we optimize the agent parameters using Adam with a learning rate of $2e - 5$. The coefficient of the objective function was $c2 = 0.5$ and $c3 = 0.05$.

2) *Task Setting*: To evaluate the performance of STTV-ASV for incremental sequence-pair speaker verification task, we simulate real-life interactive environments by configuring various parameters such as time intervals, number of daily test sessions, and sequence length. We design five scenarios to evaluate the performance, including one random scenario and four controlled scenarios.

- *Random gap limited-audio scenario*: On the first day, we randomly selected embeddings from 5 utterances as initial templates, followed by randomized day interval testing. The testing intervals ranged from 1 to 20 days, and 1 to 5 audios were randomly selected per day.
- *1 d gap all-audio scenario*: On the first day, we randomly selected 5 utterance embeddings as templates for each speaker, followed by testing every 1 day. The testing data include all audio files available on the day.

TABLE VI
RESULTS IN THE RANDOM GAP LIMITED-AUDIO SCENARIO

Method	Parameters	EER[%]	minDCF _{0.01}	
Baseline (w/o updating)	-	3.92	0.434	
FixW-TU ($\beta=0.51$)	$\alpha = 0.05$	3.12	0.363	
	$\alpha = 0.1$	2.48	0.319	
	$\alpha = 0.15$	2.20	0.310	
	$\alpha = 0.2$	2.10	0.311	
	$\alpha = 0.3$	2.16	0.315	
	$\alpha = 0.4$	2.42	0.338	
	$\alpha = 0.5$	2.73	0.354	
DRL-TU (proposed)	AdW	131k	1.99	0.283
	AdW + GT	131k	1.68	0.274
	AdW + Binary	132k	4.35	0.409
	Multi-Head (MH)	132k	1.81	0.297

+GT and +binary indicate that updating decision are determined by the ground truth label and supervised pre-trained binary action head, respectively.

- *3 d gap all-audio scenario*: Same setting as 1 d gap all-audio scenario, but the testing is conducted every 3 days.
- *5 d gap all-audio scenario*: Same as 1 d gap all-audio scenario, but the testing is conducted every 5 days.
- *10 d gap all-audio scenario*: Same as 1 d gap all-audio scenario, but the testing is conducted every 10 days.

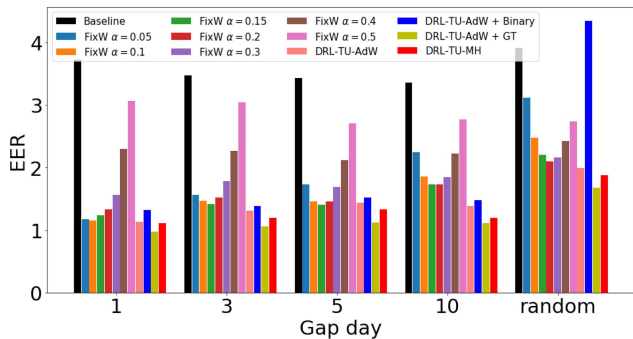
Additionally, we conduct four challenging controlled scenarios (1/3/5/10 d gap one-audio scenario), where the testing data only includes one audio per day, to evaluate the performance of DRL-based template updating. We adopt the overall EER as the primary metric, and the results of TD-EER curve and TDS curve are provided as auxiliary metrics.

B. Experimental Results and Analysis

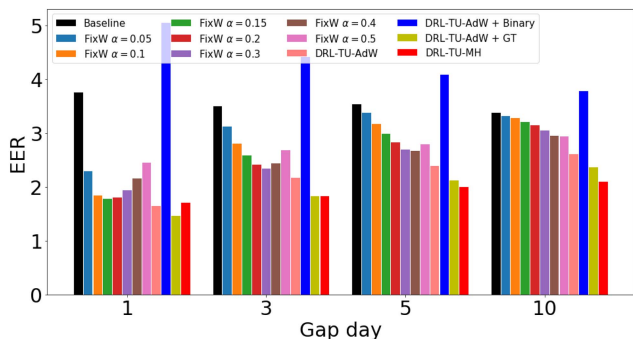
The baseline system is a one-time enrollment template without any updates. Our evaluation on the random gap limited-audio scenario in Table VI demonstrates that template updating methods improve system performance significantly compared to the baseline. In particular, the DRL-TU method outperforms the FixW-TU method due to its ability to adjust weights adaptively based on each test audio, resulting in long-term benefits.

It is worth noting that the DRL-TU method indeed uses more parameters to build the model and there are also hyper-parameters in training this DRL model as discussed in the end of Section III.C.4. For some cases where the frequency of each target user's login event is similar, it is more convenient to use the Fix-TU method as there are only two parameters to tune. However, once the DRL model is trained, it can automatically make adaptive updating and merging decisions for each testing sample which is more robust. While the updating thresholds and weights in the FixW-TU baseline are fixed for all testing samples.

Initially, we propose the DRL-TU-AdW method with a pre-set updating threshold. However, this decision-making approach often rejected positive samples and falsely accepted negative samples. Therefore, we opt to employ the ground truth (GT) as the updating decision for DRL-TU-AdW in order to investigate the impact of the updating decision. In this case, a positive sample signifies an update, while a negative sample indicates no update. The AdW-GT results show that a correct decision



(a) All-audio scenario, which refers to the scenario where all available audio samples from the day are utilized for testing.



(b) One-audio scenario, which refers to the scenario where only one audio sample from the day is utilized for testing.

Fig. 8. System performance of different template updating methods in terms of EER under various scenarios defined in Section VI-A2.

can improve the EER by up to 15% relatively. To further improve performance, we develop the DRL-TU-MH method, which addresses both updating decisions and weight prediction. We utilize the pre-trained binary action head of DRL-TU-MH for DRL-TU-AdW, replacing the threshold-based method for decision updates, which we term AdW+Binary. The results of AdW+Binary show a significant degradation in system performance. We think that the updated template is characterized by high variability, the probability of misjudgment increases with the accumulation of changes, resulting in significant performance degradation. To overcome this limitation, we utilized the multi-head PPO algorithm for fine-tuning, which ultimately led to superior performance on EER.

Moreover, we conducted a quantitative analysis of the changes in system performance at different time intervals in Fig. 8. Notably, in one-audio scenario with more than 3 days' gap, the DRL-TU-MH method is even better than GT version of DRL-TU-AdW, particularly in large time intervals as shown in Fig. 8(b). We think this situation could result from two aspects. Firstly, the weights predicted by DRL-TU-AdW may not necessarily be optimal. Secondly, even positive samples can potentially have bad cases. The combination of these two scenarios might lead to a bias in the template embedding, resulting in subpar outcomes. This implies that the update decision of DRL-TU-MH is not a binary classification but selecting meaningful test samples for updating.

In order to further evaluate the effectiveness of our proposed method, we present results of auxiliary metrics obtained from the one-day gap all-audio scenario in Fig. 9. In contrast to baseline, the trend of TD-EER and TDS^p curves for template updating methods changes slowly and gradually stabilizes over time. Although the changes in SW-EER and $SW S^p$ curves are more intense due to small sample size within each window, they also gradually stabilize as template updates accumulate. The DRL-TU based methods also achieve the best performance in the auxiliary metrics, with TDS^p curves being more stable from start to end.

Furthermore, combining the results in Figs. 8(a) and 9, we found that the performances of FixW-TU with $\alpha = 0.05/0.1/0.15$ are comparable to our proposed DRL-TU method in one-day gap all-audio scenario. However, in other scenarios, FixW-TU exhibits significantly poorer performance compared to DRL-TU based models. Therefore, we attribute the difference in performance to the **slow start-up** of FixW-TU, which requires a significant amount of accumulation to achieve optimal results.

Consequently, Fig. 10 illustrates performance under the one-day gap one-audio scenario to observe the variation trends of FixW-TU. Under the one-audio scenario, FixW-TU with a small weight needs about 10-30 days to start-up, while the FixW-TU with a large weight only needs a few days. In comparison, by comparing TDS^p (Fig. 10(b)) and $SW S^p$ (Fig. 10(d)) curves, it can be observed that DRL starts quickly and stabilizes rapidly. Furthermore, considering TD-EER (Fig. 10(a)) and SW-EER curves (Fig. 10(c)), DRL not only exhibits fast adaptation but also achieves good performance in less than 10 days and maintains it consistently. Therefore, our proposed method is a **fast response** method. Moreover, as shown in Fig. 8(b), when the number of gap days becomes larger, the advantage of FixW-TU with a large updating weight is more clear. To sum up, our proposed DRL-TU method adapts quickly and achieves the best performance in most experimental scenarios.

C. Ablation Experiments

Considering that FixW-TU requires a significant amount of manual tuning, we conducted experiments to search for the optimal values of the decision threshold (β) and weight update (α). The results are shown in Fig. 11. We performed a grid search for the update weight and threshold values within the ranges [0.1, 0.5] and [0.05, 0.9], respectively. Ultimately, we found that FixW-TU achieved the best performance when $\beta = 0.51, \alpha = 0.2$ in the random gap limited-audio scenario. In contrast, our DRL-TU-MH method can adaptively determine the updating decision which is more robust in scenarios with different or random gap days as shown in Fig. 8.

VII. DISCUSSIONS

This paper introduces novel datasets and corresponding solutions for short-term and long-term time varying speaker verification. However, the analysis of short-term time-varying scenarios reveals that the temporal changes in a speaker's voice are gradual. Therefore, solutions designed for long-term time variations are not suitable for short-term scenarios, as networks struggle

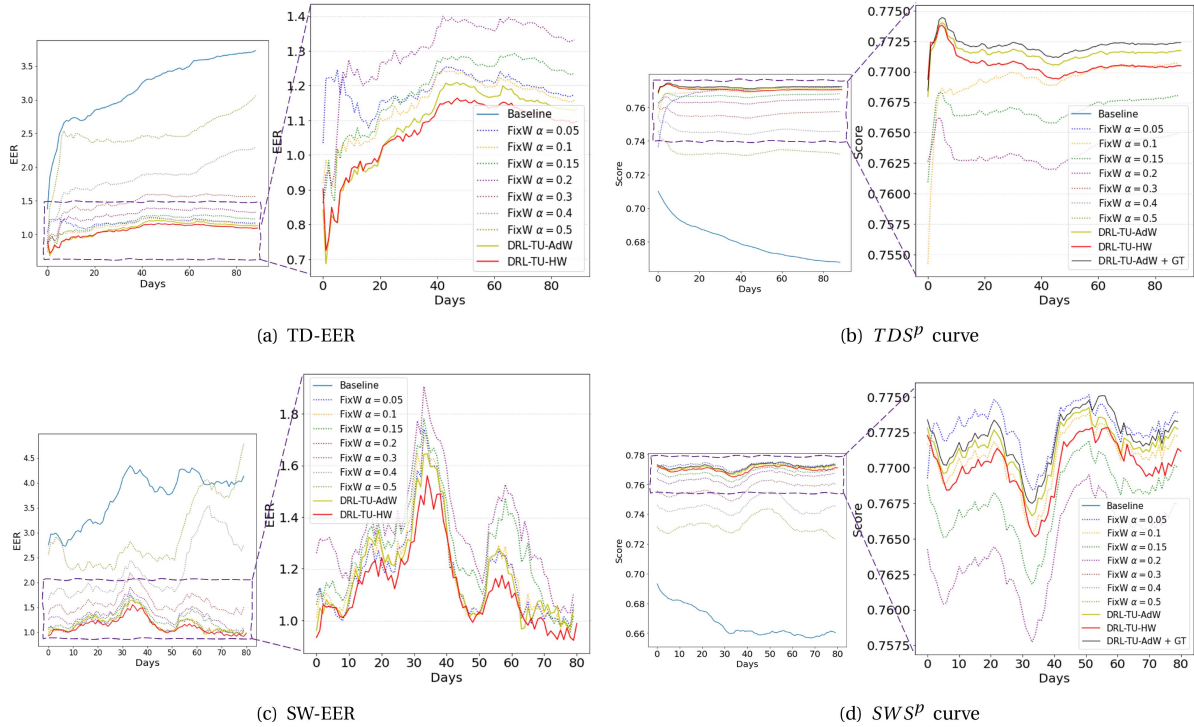


Fig. 9. System performance of different template updating methods with curves of TD-EER, TDS^p , SW-EER and SWS^p in the 1 d gap all-audio scenario defined in Section VI-A2.

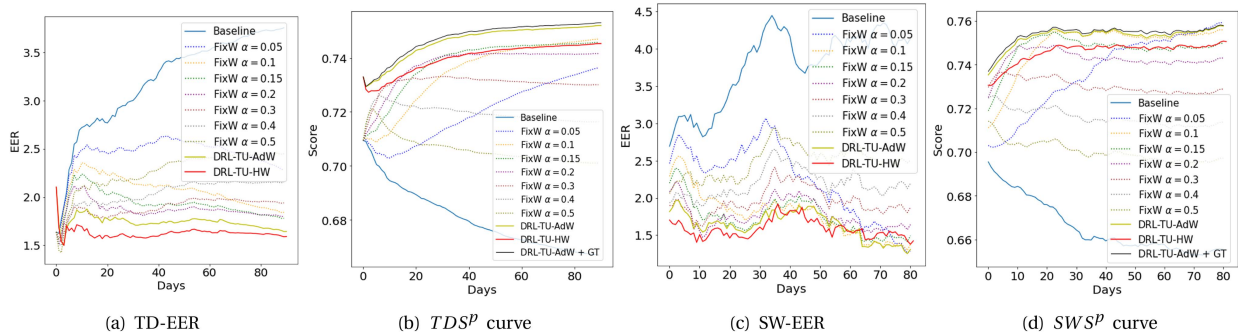


Fig. 10. System performance of different template updating methods with curves of TD-EER, TDS^p , SW-EER and SWS^p in the 1 d gap one-audio scenario defined in Section VI-A2.

to differentiate such fine-grained changes on a daily basis. On the other hand, due to the absence of a large-scale speaker recognition database continuously recorded everyday over more than a decade, applying short-term time-varying solutions to long-term scenarios is also challenging. However, if we have a continuously recorded long-term time-varying database, we believe that performance in cross-age speaker verification can also be improved through template updating or multi-template fusion methods. In future research, we will mine continuously recorded long-term time-varying data through online broadcasting channels.

There are potential issues that may arise in the practical use of DRL-TU algorithms. We think that DRL-TU-AdW ensures the stability of template updates by employing threshold

limitations but may not achieve the optimal performance. On the other hand, the DRL-TU-MH approach, where template updates are fully determined by agent, may potentially lead to better performance but with higher risks. The risk is from the potential accumulation of bad cases. If the errors increases along each template update, there is a risk of the template deteriorating over time, leading to a performance degradation in the usage. Therefore, in both FixW-TU and DRL-TU-AdW methods, we impose a minimum threshold to ensure that the updated templates do not degrade excessively. While DRL-TU-MH does not have this threshold constraint, the decision-making process is trained through supervised learning based on speaker discrimination, aiming to filter out poor positive samples.

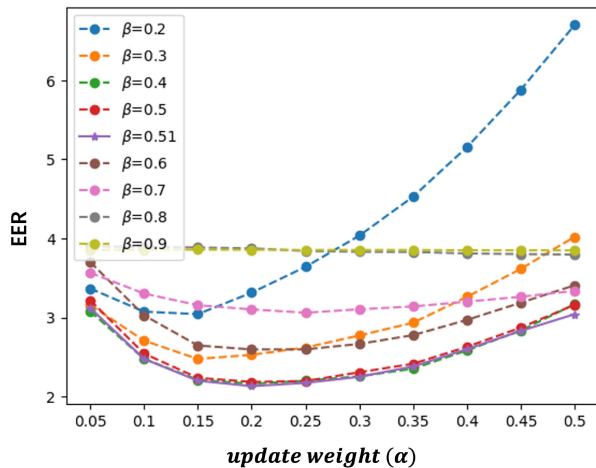


Fig. 11. Results for different β and α values in the FixW-TU under the random gap limited-audio scenario. The y-axis represents the results of EER, while the x-axis represents the values of the update weight.

Due to the limit of databases, the age distributions of Vox-Celeb and SMIP-TV are quite skewed towards young adults. However, addressing the challenge of temporal variations is not only crucial for the elderly speakers but also presents a significant challenge for adolescents going through voice changes and even children. It indeed needs a large amount of data to comprehensively study and address this issue. In our future research, we plan to enrich the age distribution of speakers, particularly focusing on the elderly people and the children, to further enhance our understanding and solutions for temporal variations across diverse age groups.

VIII. CONCLUSION

This paper proposes novel benchmarks and solutions to address the challenges of long-term and short-term time-varying speaker verification. For long-term speaker verification, we mine age information from the VoxCeleb dataset and introduce the Vox-CA test set as a benchmark for cross-age ASV tasks. Our proposed ADAL method effectively learns age-invariant speaker representation. For short-term speaker verification, we introduce the SMIP-TV dataset to investigate the challenge. We propose an incremental sequence-pair speaker verification task and adopt a template updating method to mitigate the impact of time varying. We formulate the template updating process as a Markov Decision Process and suggest a deep reinforcement learning-based method with multi-head PPO strategy to predict the updating decision and weight. Experimental results demonstrate significant improvement achieved by our DRL-TU method. Our proposed methods and released datasets contribute to robust speaker verification that can better handle time-varying scenarios.

ACKNOWLEDGMENT

Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

REFERENCES

- [1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 5329–5333.
- [2] D. 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.
- [3] D. Povey et al., "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech*, 2018, pp. 3743–3747.
- [4] 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. Speaker Lang. Recognit. Workshop Odyssey*, 2018, pp. 74–81.
- [5] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6738–6746.
- [6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4685–4694.
- [7] M. K. Nandwana, J. V. Hout, C. Richey, M. McLaren, M. A. Barrios, and A. Lawson, "The VOICES from a distance challenge 2019," in *Proc. Interspeech*, 2019, pp. 2438–2442.
- [8] X. Qin et al., "The INTERSPEECH 2020 far-field speaker verification challenge," in *Proc. Interspeech*, 2020, pp. 3456–3460.
- [9] 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 scenario," in *Proc. 2022 Far-Field Speaker Verification Challenge (FFSVC2022)*, 2022, pp. 10–14.
- [10] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV challenge 2020: Large-scale evaluation of short-duration speaker verification," in *Proc. Interspeech*, 2020, pp. 731–735.
- [11] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (SdSV) challenge 2021: The challenge evaluation plan," 2021, *arXiv:1912.06311*.
- [12] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, "Voxsrc 2021: The third voxceleb speaker recognition challenge," 2022, *arXiv:2201.04583*.
- [13] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 1985, pp. 387–390.
- [14] J.-F. Bonastre, F. Bimbot, L.-J. Boe, J. P. Campbell, D. A. Reynolds, and I. Magrin-Chagnolleau, "Person authentication by voice: A need for caution," in *Proc. Eurospeech*, 2003, pp. 33–36.
- [15] L. Wang, J. Wang, L. Li, T. F. Zheng, and F. K. Soong, "Improving speaker verification performance against long-term speaker variability," *Speech Commun.*, vol. 79, pp. 14–29, 2016.
- [16] T. Kato and T. Shimizu, "Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2003, pp. 57–60.
- [17] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space," *Comput. Speech Lang.*, vol. 27, no. 5, pp. 1068–1084, 2013.
- [18] A. Czajka, "Call for cooperation: Biometric template ageing," in *Proc. Int. Biometric Perform. Conf.*, 2010.
- [19] C. Li et al., "Deep speaker: An end-to-end neural speaker embedding system," 2017, *arXiv:1705.02304*.
- [20] A. Sholokhov, X. Liu, M. Sahidullah, and T. Kinnunen, "Baselines and protocols for household speaker recognition," in *Proc. Speaker Lang. Recognit. Workshop Odyssey*, 2022, pp. 185–192.
- [21] W. Mistretta and K. Farrell, "Model adaptation methods for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 1998, pp. 113–116.
- [22] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, 2012, pp. 478–483.
- [23] F. Kelly and N. Harte, "Effects of long-term ageing on speaker verification," in *Proc. Eur. Workshop Biometrics Identity Manage.*, 2011, pp. 113–124.
- [24] Y. Matveev, "The problem of voice template aging in speaker recognition systems," in *Proc. Speech Comput.*, 2013, pp. 345–353.
- [25] Y. Lei and J. H. Hansen, "The role of age in factor analysis for speaker identification," in *Proc. Interspeech*, 2009, pp. 2371–2374.

- [26] S. S. Xu, M.-W. Mak, K. H. Wong, H. Meng, and T. C. Y. Kwok, "Age-invariant speaker embedding for diarization of cognitive assessments," in *Proc. IEEE Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, 2021, pp. 1–5.
- [27] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 5040–5044.
- [28] P. Ghahremani et al., "End-to-end deep neural network age estimation," in *Proc. Interspeech*, 2018, pp. 277–281.
- [29] T. Gupta, D.-T. Truong, T. T. Anh, and C. E. Siong, "Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model," in *Proc. Interspeech*, 2022, pp. 1978–1982.
- [30] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Comput. Speech Lang.*, vol. 27, pp. 151–167, 2013.
- [31] P. G. Shivakumar, M. Li, V. Dhandhanian, and S. S. Narayanan, "Simplified and supervised i-vector modeling for speaker age regression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2014, pp. 4833–4837.
- [32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon, Tech. Rep. 27403, 1993.
- [33] G. Fenu, M. Marras, G. Medda, and G. Meloni, "Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition," in *Proc. Interspeech*, 2021, pp. 1892–1896.
- [34] 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 *Proc. INTERSPEECH*, 2023, pp. 1948–1952.
- [35] F. Kelly and J. H. L. Hansen, "Score-aging calibration for speaker verification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2414–2424, Dec. 2016.
- [36] N. Tawara, A. Ogawa, Y. Kitagishi, and H. Kamiyama, "Age-VOX-Celeb: Multi-modal corpus for facial and speech estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 6963–6967.
- [37] K. Hechmi, T. N. Trong, V. Hautamäki, and T. Kinnunen, "Voxceleb enrichment for age and gender recognition," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop (ASRU)*, 2021, pp. 687–693.
- [38] A. Nagrani, J. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.
- [39] T. Zheng, W. Deng, and J. Hu, "Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments," 2017, *arXiv:1708.08197*.
- [40] H. Wang, D. Gong, Z. Li, and W. Liu, "Decorrelated adversarial learning for age-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3522–3531.
- [41] Y. Wang et al., "Orthogonal deep features decomposition for age-invariant face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 738–753.
- [42] Z. Huang, J. Zhang, and H. Shan, "When age-invariant face recognition meets face age synthesis: A multi-task learning framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 7278–7287.
- [43] X. Qin, N. Li, W. Chao, D. Su, and M. Li, "Cross-age speaker verification: Learning age-invariant speaker embeddings," in *Proc. Interspeech*, 2022, pp. 1436–1440.
- [44] R. A. Cole, M. Noel, and V. Noel, "The CSLU speaker recognition corpus," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, 1998, pp. 3167–3170.
- [45] D. Li, J. Liu, Z. Wang, Y. Li, B. Chen, and L. Cai, "TRSD: A time-varying and region-changed speech database for speaker recognition," *Circuits, Syst., Signal Process.*, vol. 41, pp. 1–26, 2022.
- [46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2015, pp. 5206–5210.
- [47] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming mandarin ASR research into industrial scale," 2018, *arXiv:1808.10583*.
- [48] X. Xiang, "The xx205 system for the voxceleb speaker recognition challenge 2020," 2020, *arXiv:2011.00200*.
- [49] N. Brummer et al., "BUT+ omilia system description voxceleb speaker recognition challenge 2020," in *Proc. VoxSRC Workshop*, 2020.
- [50] S. Furui, "Recent advances in speaker recognition," *Pattern Recognit. Lett.*, vol. 18, no. 9, pp. 859–872, 1997.
- [51] F. Kelly, A. Drygajlo, and N. Harte, "Compensating for ageing and quality variation in speaker verification," in *Proc. Interspeech*, 2012, pp. 498–501.
- [52] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.
- [53] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2015, pp. 252–257.
- [54] S. Escalera et al., "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2015, pp. 243–251.
- [55] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.
- [56] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by back-propagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
- [57] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
- [58] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.
- [59] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.
- [60] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
- [61] Z. Fan, R. Su, W. Zhang, and Y. Yu, "Hybrid actor-critic reinforcement learning in parameterized action space," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 2279–2285.
- [62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
- [63] 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 Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1038–1051, 2020.
- [64] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484*.
- [65] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 5220–5224.