Speaker Verification Across Ages: Investigating Deep Speaker Embedding Sensitivity to Age Mismatch in Enrollment and Test Speech

Vishwanath Pratap Singh¹, Md Sahidullah², Tomi Kinnunen¹

¹University of Eastern Finland, Finland
²Institute for Advancing Intelligence, TCG CREST, India
vsingh@uef.fi, sahidullahmd@gmail.com, tomi.kinnunen@uef.fi

Abstract

In this paper, we study the impact of the ageing on modern deep speaker embedding based automatic speaker verification (ASV) systems. We have selected two different datasets to examine ageing on the state-of-the-art ECAPA-TDNN system. The first dataset, used for addressing short-term ageing (up to 10 years time difference between enrollment and test) under uncontrolled conditions, is VoxCeleb. The second dataset, used for addressing long-term ageing effect (up to 40 years difference) of Finnish speakers under a more controlled setup, is Longitudinal Corpus of Finnish Spoken in Helsinki (LCFSH). Our study provides new insights into the impact of speaker ageing on modern ASV systems. Specifically, we establish a quantitative measure between ageing and ASV scores. Further, our research indicates that ageing affects female English speakers to a greater degree than male English speakers, while in the case of Finnish, it has a greater impact on male speakers than female speakers.

Index Terms: ASV, ageing, gender, linear mixed effect models, VoxCeleb, ECAPA-TDNN

1. Introduction

Automatic speaker verification (ASV) systems have become increasingly important in the field of voice recognition technology including their applications in security systems, banking, call centers, law enforcement, fraud prevention, criminal investigations, and healthcare. Modern deep learning based ASV systems [1] involve three phases: (1) training of speaker embedding extractor; (2) enrollment to create a reference model of the target speakers; and (3) verification to validate a speaker’s claimed identity based on the similarity between the enrollment and verification recordings (ASV score).

Similar to other biometrics, ASV systems are not perfect. It is evident that the accuracy of the ASV system depends strongly on how closely the enrollment and verification conditions are matched [2, 3]. Over the years, extensive prior research works have addressed the effects of various mismatch factors between enrollment and verification phases, including background environment [4], speaker-to-microphone distance [5], language [6], and speaking style [2, 7, 8]. In this study, we focus on another fundamental mismatch factor that has received somewhat less attention: speaker ageing due to the time lapse between the enrollment and verification recordings.

ASV systems utilize distinctive acoustic characteristics of an individual’s voice to authenticate his or her identity. However, the impact of ageing on the human voice may alter these acoustic characteristics, which in turn can impact the performance of speaker verification systems. Specifically, as a person ages, their vocal tract, vocal folds, and related organs undergo changes that result in variations in measurable acoustic parameters, including pitch, formants [13], and loudness [14]. Furthermore, the impact of the speaker ageing may depend substantially on speaker’s gender and articulatory behavior [15]. Hence, a holistic study of ageing impact on the speaker verification systems should also consider the factors such as gender and language.

Despite having contributed to improved understanding of voice ageing and its impact on ASV systems, the prior studies have a number of limitations. In particular, the existing studies are limited by smallish sample sizes, controlled environments [16, 17], short-term ageing [18], limited number of age groups and sessions [19]. In addition, they have focused on classical GMM-UBM [20] and i-vector based ASV systems [16, 17, 21]. In this paper, we have studied the impact of ageing on ASV scores across gender, datasets, languages, and different session conditions with ECAPA-TDNN [22] based ASV system.

The selection of appropriate datasets is vital for addressing the shortcomings of prior studies on age-related effects. In this study, we consider two datasets that are vastly different. The first dataset is VoxCeleb [9, 23], a commonly used dataset to benchmark ASV systems. In order to examine the impact of ageing on English speech, we have restricted our analysis to speakers from the USA. The original VoxCeleb dataset, however, does not contain information about the age of speakers. In this study, we rely on the estimated age information in Vox-
Table 1: Age and Gender Enriched VoxCeleb1 and Longitudinal Corpus of Finnish Spoken in Helsinki (LCFSH) Details

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoxCeleb1 Official [9]</td>
<td># Speakers 98350</td>
<td># Speakers 89733</td>
</tr>
<tr>
<td>VoxCeleb1 Enriched [10]</td>
<td># Speakers 561</td>
<td># Speakers 9792</td>
</tr>
<tr>
<td>VoxCeleb1-Age-Enriched (USA)</td>
<td># Sessions 2.70</td>
<td># Sessions 3.66</td>
</tr>
<tr>
<td>LCFSH [11]</td>
<td># Utterances 80960</td>
<td># Utterances 9740</td>
</tr>
</tbody>
</table>

Table 2: Verification Accuracy (%EER) Across Gender on VoxCeleb1-Age-Enriched (USA). Here, session condition Both includes recordings from both same and different sessions in trial pairs, while session condition Different excludes recordings from the same session in trial pairs.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age Difference Between Enrolment and Test Pairs (Years)</th>
<th>Verification Accuracy (%EER) Across Gender on VoxCeleb1-Age-Enriched (USA)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 (Both)</td>
<td>0 (Different)</td>
</tr>
<tr>
<td>Male</td>
<td>1/2/3/4/5/6/7/8/9/10</td>
<td>1/2/3/4/5/6/7/8/9/10</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>1/2/3/4/5/6/7/8/9/10</td>
</tr>
</tbody>
</table>

Table 3: Verification Accuracy (%EER) Across Gender on Longitudinal Corpus of Finnish Spoken in Helsinki (LCFSH) Dataset.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age Difference (Years)</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-10</td>
<td>5.12</td>
<td>10.04</td>
</tr>
</tbody>
</table>

To summarize, prior works on the impact of ageing on ASV

| 2. Related Works |

Prior research examining the effects of the ageing on ASV can be divided broadly into two main categories: (1) investigating change in ASV scores; and (2) analyzing and mitigating the impact of ageing on ASV accuracy. We briefly review both kinds of work below.

2.1. Analysing the Impact of Ageing on ASV Score

In [16, 17], authors created a new dataset for studying age-related changes in ASV, comprising 9 male and 9 female speakers drawn from the British Broadcasting Corporation (BBC) and Raidió Teilifís Éireann (RTE) broadcasts. The main findings from these two studies are: (1) the effect of the ageing is more pronounced on the target (same-speaker) trial scores compared to the non-target (different-speaker) trial scores; and (2) the mean target score gets lowered due to the ageing, yielding increased equal error rates (EERs). Moreover, this study was conducted in controlled recording environments with a small sample size of speakers with classical GMM-UBM [20] based ASV systems. In another work [18], the study of short-term ageing on the GMM-UBM based ASV system was conducted with age differences up to three years. The authors concluded that the impact on the ASV due to the short-term ageing (0-3 years) was insignificant compared to the session variability. Similar to previous studies [16, 17], the analysis is conducted using Multi-Session Audio Research Project (MARP) corpus with a limited number of speakers, sessions, and in a gender independent manner. In [19], only two age groups, one between 20-30 years of age (‘young’) and another between 40-50 years of age (‘old’), were included for studying the effect of absolute age on factor analysis for speaker recognition.

2.2. Analyzing and Mitigating the Impact of Ageing on ASV Accuracy

Besides the above studies, a number of studies have also addressed the impact of ageing on verification accuracy and proposed improvements to ASV systems in this regard. The impact of age difference between target and non-target speakers on speaker recognition performance has been investigated separately for male and female speakers [25]. In another study [21], the authors proposed the score calibration to mitigate the impact of the ageing on a i-vector based ASV system. Recently, an approach for age-invariant speaker embedding was studied on the VoxCeleb dataset [26]. The primary focus of this study was to assess the improvement in the EER metric.

To summarize, prior works on the impact of ageing on ASV systems have typically used small sample sizes. These studies have utilized conventional GMM-UBM and i-vector-based ASV systems. Furthermore, previous studies have focused primarily on the English language, with limited attention to other languages.
3. Experimental Setup

3.1. Datasets

In order to train the speaker embedding extractor, we have utilized the development portion of the VoxCeleb2 dataset [23], which consists of 5,994 speakers. We randomly split the data into a 90:10 ratio for training and validation. Further, five additional samples for each utterance are generated using 2-fold time domain SpecAugment [27] and 3-fold augmentations with combinations of publicly available MUSAN dataset (babble-noise) [28] and the room impulse response (RIR) dataset (reverb) [29].

In the first part of the analysis, we have used the VoxCeleb1 dataset [9] for studying the impact of the ageing. We have utilized the gender, geography (country), and age metadata from [10]. To minimize the influence of geographic location, accent, and language differences we specifically chose to include only those sessions from VoxCeleb1 featuring speakers from the USA, the most represented country in the VoxCeleb1 dataset. We refer to this gender and age enriched dataset as VoxCeleb1-Age-Enriched (USA) in the rest of the paper. Distributions of the number of speakers and sessions in the different evaluation protocols of VoxCeleb1-Age-Enriched(USA) are presented in Table 1. Further, Fig. 4 illustrates the distribution of speakers in the VoxCeleb1-Age-Enriched (USA) dataset, categorized by gender.

In the second part, we have conducted experiments with the Longitudinal Corpus of Finnish Spoken in Helsinki (LCFSH) [11] dataset to study the impact the ageing in Finnish. The LCFSH dataset is collected through interviews with the native residents of Helsinki with varying social backgrounds. The dataset contains a total of 154 speakers of which 56 speak native residents of Helsinki with varying social backgrounds. The LCFSH dataset is collected through interviews with the Longitudinal Corpus of Finnish Spoken in Helsinki (LCFSH) [11] dataset to study the impact the ageing in Finnish.

The dataset contains 154 speakers of which 56 speak native residents of Helsinki with varying social backgrounds. The LCFSH dataset is collected through interviews with the Longitudinal Corpus of Finnish Spoken in Helsinki (LCFSH) [11] dataset to study the impact the ageing in Finnish.

3.2. Speaker Verification System

We conduct all the model training and evaluation experiments using the SpeechBrain toolkit [30]. Specifically, we use the ECAPA-TDNN based ASV system that achieves state-of-the-art performance [22]. We use cosine similarity measure followed by the score normalized using adaptive s-norm [31, 32].

The ECAPA-TDNN model trained on VoxCeleb2 achieves EER of 1.28% and 1.95% on the Vox1-O and Vox1-E datasets, respectively.

4. Methodology

We use the linear mixed effect (LME) [24] models to analyze the variation in the ASV score with respect to the age difference under different gender, language, and session conditions. We fit the following model to ASV score data under different conditions:

\[ S_i = \mu + \beta \cdot \delta_i + \gamma \]  

where \( S_i \) is ASV score of \( i \)-th trial, \( \mu \) is intercept or mean of the ASV scores in a particular condition, \( \delta_i \) is the age difference between enrollment and verification audio, \( \beta \) is fixed effect coefficient that establishes a linear relationship between age difference (\( \delta_i \)) and ASV score (\( S_i \)), and \( \gamma \) is a random effect which is induced by other variabilities such as background noise, recording setup, etc. We use the Statsmodels module in Python for the LME analysis [12].

In the previous study [15], it is evident that the ageing impacts male and female speakers differently. Hence, we have conducted the analysis using different LMEs across different gender conditions. Further, the effect of the ageing on the ASV scores varies significantly across target and non-target scores [16, 17]. Therefore, we also conduct separate analyses for target and non-target trial conditions. Finally, target trials from the same session can result in higher levels of similarity, which can introduce variations in ASV scores and create confusion with variations caused by the ageing. For this reason, we conduct separate analysis for different session conditions as well.
Table 4 presents a total of six separate cases based on different gender, trial, and session conditions. When selecting target trial pairs with an age difference of zero in the VoxCeleb1-Age-Enriched (USA) dataset, the pairs may originate from the same session or different sessions. For this study, we have taken into account both scenarios, including cases where trial pairs from the same session are present or excluded when there is no age difference between them. We represent these two session conditions as Both and Different in Table 2 and 4. On the other hand, LCFSH corpus contains a single session for each speaker and hence the second session condition Different is not applicable for this corpus and consequently C-2 and C-5 are not studied in Table 4.

## 5. Results and Discussions

### 5.1. Analyzing the Impact of Ageing on VoxCeleb1-Age-Enriched (USA)

Figure 4a displays the gender and age distribution of speakers in VoxCeleb1-Age-Enriched (USA), while Table 2 presents the corresponding EERs. We observe in Table 2 that trends of degradation in EERs with respect to the ageing are not consistent. However, when observing the EERs at three-year intervals (i.e., EERs for age gaps of 0, 3, 6, and 9 years), we observe a monotonic increase in EERs. One possible explanation could be that the within-speaker variability is overshadowing the very short-term (0-2 years) ageing mismatches.

Further, changes in intercepts of C1 vs. C2 and C4 vs. C5 in Table 4, as well as the plots in Fig. 2a vs. Fig. 2b and Fig. 2c vs. Fig. 2d indicate that the predicted target LME score is in a higher range when target trials from the same session are present. This suggests that the reason for a high LME score could be attributed to the similarity of the recording conditions or environment, particularly when trial pairs are chosen from the same session.

We also observe from the LME parameters in Table 4 that the fixed effect coefficients are negative in all the conditions (C1-6). This indicates both male and female speakers have a negative impact on ASV scores due to the ageing. In other words, as a consequence of ageing, the similarities between embeddings decrease for all types of trials across different conditions. Further, we observe that non-target scores (C3 and C6) have low fixed effect coefficients which are considerably smaller in magnitude than that of target scores (C1, C2 and C4, C5). This indicates that the target scores are more sensitive to the speaker ageing compared non-target scores.

Additionally, by examining the parameters of C1 and C4 on the VoxCeleb1-Age-Enriched (USA) dataset in Table 4, we observe that the fixed effect coefficient of C4 is comparatively greater than that of C1. This suggests that ageing might have a stronger influence on male target scores than on male target scores in the English language.

### 5.2. Analyzing the Impact of Ageing on LCFSH

The EERs reported in Table 3 for the LCFSH dataset is significantly higher than the EERs reported in Table 2 for VoxCeleb1-Age-Enriched (USA). This might be due to the domain mismatch between train and evaluation conditions between VoxCeleb and LCFSH induced by the differences in language, recording set-up, background environment, etc. Moreover, the degradation in EER within the LCFSH dataset for different ageing conditions is consistent and well-evident, likely due to age groups with larger age gaps.

Similar to VoxCeleb1-Age-Enriched (USA), all fixed effect coefficients (C1, C3, C4, and C6) are negative in the LCFSH dataset. The fixed effect coefficients for target scores (C1 and C4) are notably larger in magnitude than those for non-target scores, indicating that, like in English, Finnish target scores are more sensitive to ageing than non-target scores. Furthermore, we observe in Fig. 3b and Fig. 3d that the x-axes for non-target scores are limited to a narrow range, while the target scores in Fig. 3a and Fig. 3c are distributed more widely. This further confirms that target scores are more impacted due to ageing.

Unlike the results from VoxCeleb1-Age-Enriched (USA), the fixed effect coefficient for C4 is smaller than that for C1 in the LCFSH dataset. This suggests that, in Finnish, female target scores are less impacted than male target scores. Hence, we conclude that ageing impacts male and female speakers differently and the impact is not consistent across the languages. Furthermore, overall fixed effect coefficients of LME models trained on LCFSH dataset are smaller in magnitude than the corresponding LME models trained on VoxCeleb1-Age-Enriched (USA) dataset. This indicates that the ageing has less impact on Finnish compared to the English. This might be due to the differences in articulation styles [15] between the two languages.

## 6. Conclusions

We examined the impact of ageing on the modern ECAPA-TDNN based speaker verification system. Using two publicly available datasets, we investigated the impact of ageing using a linear mixed effect model for different genders, session conditions, and languages. Our study reveals that the ageing impacts male and female speakers differently and the impact is dependent on the language. Considering trials from different sessions, we found that target scores are more dependent on age differences than non-target scores. This study utilized the estimated age for the speakers of the VoxCeleb dataset where the actual age might be different. This calls for further investigation of the issue with more accurate age information for this large dataset. The proposed methodology can also be useful for assessing age-invariant ASV systems.

## 7. Acknowledgment

This work was partially supported by Academy of Finland (Decision No. 349605, project “SPEECHFAKES”)
8. References


