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

Vishwanath Pratap Singh[1], Md Sahidullah[2], Tomi Kinnunen[1]

[1]University of Eastern Finland, Finland
[2]Institute 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 enrolment 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

---

A GitHub repository containing all the metadata and scripts will be published after the Interspeech decision.
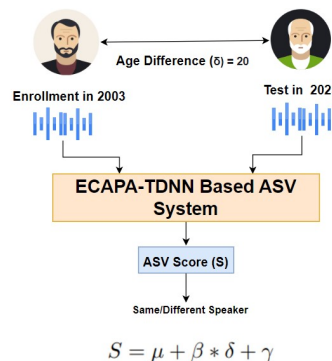

Figure 1: *We attempt to explain the degraded ASV score (S) caused by the speaker ageing, with respect to age difference (δ) across gender and different language using two linear mixed effect model (LME) parameters namely, fixed effect due to the ageing (β) and random effect (γ) due to the various other factors across the sessions.*

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*

Datasets	Male			Female		
	# Speakers	# Sessions	# Utterances	# Speakers	# Sessions	# Utterances
VoxCeleb1 Official [9]	690	12704	90450	561	9792	63066
VoxCeleb1 Enriched [10]	678	12476	88873	536	9357	60306
VoxCeleb1-Age-Enriched (USA)	371	6584	46760	299	5081	32303
LCFSH [11]	51	51	7740	58	58	7734

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.*

Gender	Age Difference Between Enrolment and Test Pairs (Years)											
	0 (Both)	0 (Different)	1	2	3	4	5	6	7	8	9	10
Male	1.25	1.60	2.30	2.18	2.31	2.38	2.98	2.84	2.66	3.07	3.63	3.49
Female	1.34	1.59	2.29	3.91	2.70	2.70	3.05	3.12	3.66	3.41	3.87	3.89

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

Gender	Age Difference (Years)		
	0	20	40
Male	5.12	18.68	51.61
Female	10.04	21.67	40.61

Celeb provided in [10].

Our second dataset is *Longitudinal Corpus of Finnish Spoken in Helsinki* (LCFSH) [11], collected between 1970 to 2013 with consistent intervals of approximately 20 years. Unlike the VoxCeleb dataset which includes a mixture of different languages, accents, and recording conditions, the LCFSH dataset is exclusively in the Finnish language. Moreover, this is a controlled dataset with no uncertainty of the speaker's age. Hence, it is a good choice for studying the *long-term* ageing effect on languages besides English.

Moreover, the impact of ageing on the ASV systems may have a combination of *fixed effects* due to the age difference as well as *random effects* due to various other factors across sessions. Hence, it is equally essential to consider utilizing appropriate analysis methods, such as *linear mixed effect* (LME) [24] models, which enable jointly modeling the random and fixed effects in ASV scores. LMEs are a type of statistical model used to analyze *grouped data* using regression techniques.

To summarize, our study has the following main contributions:

- Unlike most prior studies that have analyzed the ageing effect using a single language only, our study includes two very different languages and datasets namely, VoxCeleb and LCFSH.

- Studies are conducted separately for male and female speakers to understand the impact of ageing across the gender.

- As opposed to the previous studies that rely mean of scores, we have conducted extensive experiments using the linear mixed effect (LME) models to analyze the fixed and random effect on the ASV scores due to the ageing and other variations, such as recording setup and background noise.

- Unlike the existing studies which use traditional GMM-UBM or i-vector system, we perform all the experiments with state-of-the-art ECAPA-TDNN [22] based ASV systems.

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 ageing, comprising 9 male and 9 female speakers drawn from the British Broadcasting Corporation (BBC) and $Raidiò\ Teilifìs\ Èireann$ (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.
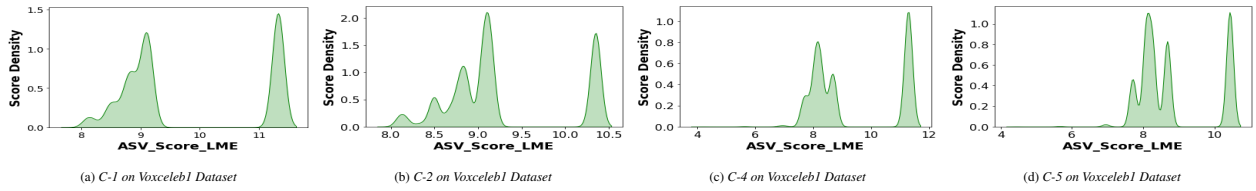
(a) C-1 on Voxceleb1 Dataset   (b) C-2 on Voxceleb1 Dataset   (c) C-4 on Voxceleb1 Dataset   (d) C-5 on Voxceleb1 Dataset

Figure 2: *Prediction of the ASV score density from LME on VoxCeleb1-Age-Enriched(USA) dataset under different conditions. The Statsmodels module in Python enables score density prediction using the LME model [12]. Due to the space constraint, we have not plotted the non-target scores.*
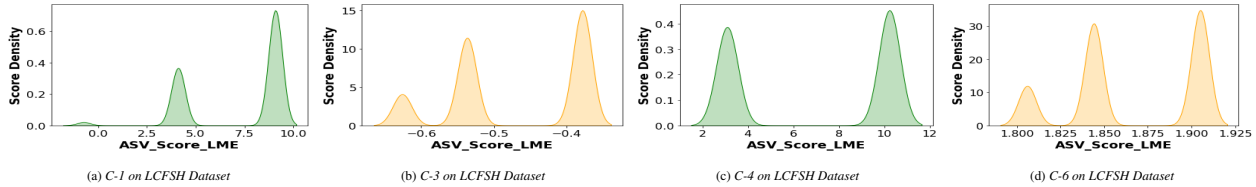

(a) C-1 on LCFSH Dataset   (b) C-3 on LCFSH Dataset   (c) C-4 on LCFSH Dataset   (d) C-6 on LCFSH Dataset

Figure 3: *Prediction of the ASV score density from LME on LCFSH dataset under different conditions. The green plots are corresponding to target scores while orange plots are corresponding to non-target scores.*
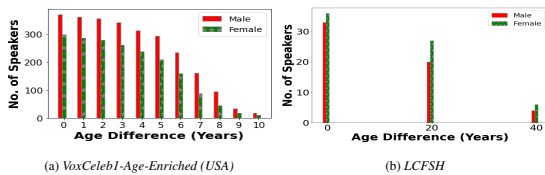

(a) VoxCeleb1-Age-Enriched (USA)   (b) LCFSH

Figure 4: *Distribution of speakers with respect to age difference in (a.) VoxCeleb1-Age-Enriched (USA), and (b.) LCFSH dataset.*

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 Vox-Celeb1 dataset [9] for studying the impact of the ageing. We have utilized the gender, geography (country), and age meta-data 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 speakers have sessions with an age difference of approximately 20

years, and 13 speakers have sessions with an age difference of approximately 40 years. Number of speaker for each gender and age difference are illustrated in Fig. 4. Number of speakers, sessions, and utterances available after the pre-processing are shown in Table 1.

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:

$$S_i = \mu + \beta * \delta_i + \gamma \tag{1}$$

where S_i is ASV score of i-th trial, μ is intercept or mean of the ASV scores in a particular condition, δ_i is the age difference between enrollment and verification audio, β is fixed effect coefficient that establishes a linear relationship between age difference (δ_i) and ASV score S_i, and γ 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: *Various conditions for LME analysis and corresponding LME parameters. Here, session condition* Both *includes recordings from both the same and different sessions in trial pairs, while session condition* Different *includes recordings only from the different sessions in trial pairs. C1-C6 in Fig. 3 and Table 4 correspond each other.*

Conditions	Gender	Score	Session	VoxCeleb1-Age-Enriched (USA)			LCFSH		
				Intercept	Fixed Effect	Random Effect	Intercept	Fixed Effect	Random Effect
C1	Male	Target	Both	9.74	-0.164	0.493	8.664	-0.246	0.029
C2	Male	Target	Different	9.453	-0.128	0.217	-	-	-
C3	Male	Non-target	Different	-1.85	-0.014	0.001	-0.869	-0.002	0.028
C4	Female	Target	Both	9.933	-0.399	0.924	8.687	-0.183	0.036
C5	Female	Target	Different	9.649	-0.357	0.669	-	-	-
C6	Female	Non-target	Different	-1.896	-0.014	0.001	1.471	-0.001	0.005

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 represents 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 *v.s.* C2 and C4 *v.s.* C5 in Table 4, as well as the plots in Fig. 2a *v.s.* Fig. 2b and Fig. 2c *v.s.* 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 aging, 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 female 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

8. References

[1] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[2] A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination performance for "easy" versus "hard" voices in style-matched and -mismatched speech," *Journal of Acoustical Society of America*, vol. 151, no. 2, p. 1393, Feb. 2022.

[3] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. Reynolds, "The NIST speaker recognition evaluation-overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2, pp. 225–254, 2000.

[4] G. Pradhan, H. B C, S. Prasanna, and R. Sinha, "Speaker verification in sensor and acoustic environment mismatch conditions," *International Journal of Speech Technology*, vol. 15, pp. 381–392, 09 2012.

[5] X. Qin, M. Li, H. Bu, W. Rao, R. K. Das, S. Narayanan, and H. Li, "The INTERSPEECH 2020 far-field speaker verification challenge," in *Proc. Interspeech*, 2020, pp. 3456–3460.

[6] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5816–5820.

[7] R. González-Hautamäki, M. Sahidullah, V. Hautamäki, and T. Kinnunen, "Acoustical and perceptual study of voice disguise by age modification in speaker verification," *Speech Communication*, vol. 95, pp. 1–15, 2017.

[8] R. González Hautamäki, V. Hautamäki, and T. Kinnunen, "On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise," *Journal of Acoustical Society of America*, vol. 146, no. 1, p. 693, Jul. 2019.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017.

[10] K. Hechmi, T. N. Trong, V. Hautamäki, and T. Kinnunen, "Voxceleb enrichment for age and gender recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 687–693.

[11] I. O. Bartis, *The longitudinal corpus of Finnish spoken in Helsinki (1970s, 1990s and 2010s)*. FIN-CLARIN-konsortio, Nykykielten laitos, Helsingin yliopisto, 2015.

[12] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[13] U. Reubold, J. Harrington, and F. Kleber, "Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers," *Speech Communication*, vol. 52, no. 7-8, pp. 638–651, Jul. 2010.

[14] P. B. Mueller, "The aging voice," *Seminars in Speech and Language*, vol. 18, no. 2, pp. 159–68; quiz 168–9, May 1997.

[15] S. Taylor, C. Dromey, S. Nissen, K. Tanner, D. Eggett, and K. Corbin-Lewis, "Age-related changes in speech and voice: Spectral and cepstral measures," *Journal of Speech, Language, and Hearing Research : JSLHR*, vol. 63, pp. 1–14, 02 2020.

[16] F. Kelly and N. Harte, "Speaker verification with long term ageing data," in *5th IAPR International Conference on Biometrics (ICB)*, 2012, pp. 478–483.

[17] F. Kelly and H. Naomi, "Effects of long-term ageing on speaker verification," *Biometrics and ID Management*, pp. 113–124, 2011.

[18] A. D. Lawson, A. R. Stauffer, B. Y. Smolenski, B. B. Pokines, M. Leonard, and E. J. Cupples, "Long term examination of intra-session and inter-session speaker variability," in *Interspeech 2009*. ISCA: ISCA, Sep. 2009.

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

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

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

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

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, pp. 1–48, 06 2015.

[25] G. Doddington, "The effect of target/non-target age difference on speaker recognition performance," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2012)*, 2012, pp. 263–267.

[26] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Cross-age speaker verification: Learning age-invariant speaker embeddings," in *Proc. Interspeech*, 2022.

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[28] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[29] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[30] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[31] Z. N. Karam, W. M. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4512–4515.

[32] S. Cumani, P. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," 08 2011, pp. 2365–2368.