Voice or Content? — Exploring Impact of Speech Content on Age Estimation from Voice

Yuta Ide*, Naohiro Tawara†, Susumu Saito*, Teppei Nakano* Tetsuji Ogawa*
*Dept. of Communications and Computer Engineering, Waseda University, Tokyo, Japan,
†NTT Communication Science Laboratories, Kyoto, Japan

***Abstract*—To investigate the impact of speech content on age estimation accuracy from voice data, we created a corpus of speech utterances featuring identical content spoken by individuals of varying ages. Subsequently, we analyzed the age estimation outcomes derived from this dataset. Previous studies have identified biases in age-labeled speech corpora regarding speaker age and vocabulary usage. Given that speech content typically varies with the speaker's age during conversations, it's plausible that age estimation results could be influenced by speech content. To address this concern, we developed a dataset in which speakers of different ages delivered speech content that was consistent across all speakers and tailored to the characteristics of each age group. We estimated the speakers' ages both manually, through crowdsourcing, and automatically and then conducted a significance test to assess whether speech content affected the age estimation results. Our findings indicated that neither the automatic age estimation system employed nor the manual age estimation outcomes were significantly impacted by speech content.**

***Index Terms*—Statistical hypothesis testing, crowdsourcing, age estimation, speaker attribute estimation**

I. Introduction

The ability to estimate attributes such as a speaker's age and emotions solely from their voice has the potential to enhance spoken dialogues and enrich the user experience in call centers. Recent advancements in deep learning techniques [1], [2] and the availability of large, age-labeled datasets [3] have significantly improved the accuracy of automatic age estimation from voice. Moreover, age information can also enhance the recognition performance of other attributes. For instance, studies have demonstrated that explicitly considering a speaker's age can improve the accuracy of speaker identification [4].

Age estimation based on the voice has been extensively studied, with research focusing on vowels [5], [6], read speech [7]–[9], and spontaneous speech [10]. These studies have shown a strong correlation between the speaker's actual age and the age estimated manually [11]. However, such experiments suffer from limitations in sample diversity and attributes, and there is a dearth of large-scale studies in this field. Furthermore, it has been noted that age labels in large age-labeled datasets may be biased [12]. Studies on system-based age estimation have revealed the existence of age-based vocabulary bias in a public corpus of spontaneous speech [13], with training the estimator with this bias leading to improved age estimation performance. Interestingly, research has also shown that the language spoken by the speaker and the listener can impact manual age estimation performance [14]. Additionally, speech content in everyday conversations changes with age, which suggests that age estimation results may be influenced by the content of the speech rather than just the acoustic information. However, the impact of speech content on age estimation in both automatic age estimation systems and manual age estimation has not been thoroughly investigated.

Our study, therefore, aims to explore the impact of speech content on age estimation, specifically investigating whether the age of speakers can be accurately estimated based solely on acoustic characteristics or if speech content plays a role in age estimation. For instance, when a young person and an elderly person discuss a topic typically associated with youth, we aim to clarify whether age estimation is influenced by the content of the speech, resulting in both speakers being perceived as young, regardless of their actual age. To achieve our objective, we require a dataset consisting of speakers of different ages speaking the same sentence, with each speaker's age appropriately labeled. Unfortunately, such a publicly available dataset is not currently accessible. As an alternative, we select sentences that are characteristic of younger and older individuals and compiled an audio dataset with the help of multiple speakers from diverse age groups to read these sentences out loud through crowdsourcing, to investigate the performance of age estimation systems and manual age estimation. Statistical significance tests are conducted between the estimated ages of two sentences to clarify whether speech content can impact age estimation results.

This study provides the following contributions:

1) The development of a speech corpus comprising the same content spoken by individuals of varying ages.
2) A method for selecting age-biased speech content, e.g., content typically spoken by younger and older individuals.
3) Insight into the impact of voice and speech content on age estimation.

The results of this study will be beneficial for researchers and engineers working on the design and development of age estimation systems from voice data.

The rest of the present paper is organized as follows: Section II introduces Tutti, an engine software developed to facilitate the use of a crowdsourcing framework for voice recording and subjective evaluation. Section III outlines the voice recording method employed in our study, detailing the process of selecting natural sentences as speech content for

individuals from younger and older age groups and recording voices while maintaining the naturalness of speech. Section IV discusses the age estimation system and the manual age estimation method used in this study. Section V presents the results of the analysis, and Section VI provides a summary of this paper.

II. Engine Software for Utilizing Crowdsourcing

For the audio recording and subjective evaluation in this study, we leveraged crowdsourcing, with a specific focus on Amazon Mechanical Turk (MTurk). Here, we will discuss the engine software that we created to streamline the use of crowdsourcing [15], [16].

We used our originally developed Tutti to design a microtask user interface (UI) for crowdsourcing. Tutti is an engine software that makes it easy to design a web UI for outsourcing annotation work as microtasks. When using crowdsourcing to perform large-scale annotation work, in many cases, a large number of system implementations are required, such as a mechanism to distribute different data on the same UI and a mechanism to collect responses from many workers. This requires a large amount of time to complete the experiment. On the other hand, Tutti obviates the need to be concerned with anything beyond the data to be annotated or evaluated, thereby substantially reducing the time invested in the experiment.

The main advantages of using Tutti in this experiment are:

- A web page template with the ability to load different data, which allows for quick preparation of data collection by making only a few UI design changes and uploading the data to be loaded.
- The ability to design transition diagrams for multiple types of web pages, enabling users to design complex tasks that require repetitive labeling of the same UI within the same micro-task or provide different UIs for varying conditions.
- The function of automatically assigning tasks to workers, making it possible to outsource micro-tasks appropriately according to the target number of responses to be collected for each presented data.
- Immediate confirmation of collected worker responses on the GUI console or through the API.

III. Voice Recording

Younger and older individuals may use different vocabulary and discuss distinct topics. However, for instance, when a young person and an elderly person discuss the same topics that a young person typically talks about, it may not be clear if the age of both speakers can be accurately estimated based on the acoustic characteristics of their speech or if the content of their speech affects the estimation, leading to the speaker being perceived as young, regardless of their actual age.

To explore the impact of speech content on age estimation, we chose sentences that are typically spoken by younger and older age groups. We then asked several individuals of varying ages to read the selected sentences aloud as naturally as possible, and recorded their speech. We conducted the voice recordings through a crowdsourcing task on MTurk.

In this section, we explain the process of selecting appropriate sentences for younger and older individuals, as well as provide guidance on recording speech that sounds natural.

A. Selecting Sentences with Content Biased Towards Particular Age Group

To investigate whether speech content affects age estimation by voice, we chose sentences from AgeVoxCeleb [3], a speech corpus labeled with age, that are typically spoken by younger or older people. We then had these sentences recorded by crowdworkers. Our approach for selecting these sentences is detailed in the rest of this section.

Our method for selecting sentences involved choosing English utterances from AgeVoxCeleb where the speaker was either under 30 or over 60 years old. We then used speech recognition to obtain the corresponding sentences (as a series of characters). To generate the character series, we used a simple greedy method on a sequence of acoustic feature vectors extracted through wav2vec 2.0 [17]. We obtained a feature representation of the sentences using sentenceBERT [18]. Next, we clustered the representations into two groups using the k-means algorithm and selected the 20 sentences closest to the centroid of each cluster. We found a statistically significant difference between the two clusters based on the actual age of the speaker for each of the 20 sentences in each cluster. We labeled the cluster with the higher average age of the speaker as the "older" cluster and the other as the "younger" cluster. Finally, we chose five sentences from the "older" cluster where the speaker was 60 years old or older, and five sentences from the "younger" cluster where the speaker was under 30 years old based on their proximity to the centroid of each cluster.

To enable workers to speak as naturally as possible, we modified the selected sentences. The speech recorded in AgeVoxCeleb consists of shortened versions of press conferences or speeches of a certain length. Consequently, it might be challenging for workers to speak spontaneously if the sentences are used as is, due to the lack of subjects or grammatical errors. To address this issue, we modified the selected sentences as much as necessary to allow for natural speech while ensuring that the meaning of the sentences remained unchanged by referring to the original press conferences and speeches. Table I presents an example of the corrections made.

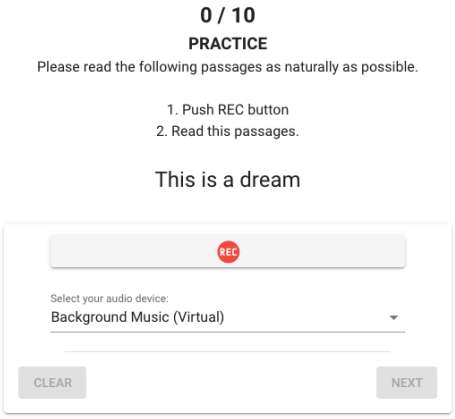
B. Recording Voice with Naturalness of Speech

Simply reading the sentences aloud could result in a loss of natural speech, which could potentially impact the accuracy of age estimation results. This is less than ideal for a survey. To capture the most natural speech possible, we implemented a speaker check after each recording to ensure the naturalness of the speech. We also used the final recorded speech after multiple recordings to achieve the best possible results.

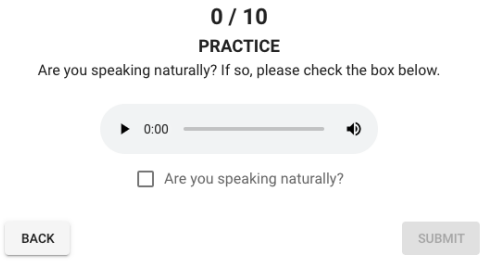
Figure 1 displays the voice recording interface used by crowdworkers. During the recording phase (Fig. 1(a)), a sentence is displayed on the screen and the workers record their

TABLE I: Example of sentence revision. Text was revised to be natural, taking care not to change meaning of sentence.

Before correction	After correction
a very good match actually had a a good start up to five love and nem she find a way back into the math and em made it very difficult for me and em ye	It was a very good match. Actually, I did a good start-up to five love, and then she found a way back into the match and made it very difficult for me.
at a period of time were with us change that's taking place that we have to start making some	We're at a period. We're with this change that's taking place. We have to start making some strategic decisions.



(a) User interface for recording phase.



(b) User interface for confirmation phase.

Fig. 1: Screen where workers perform recording process. Workers speak to capture their speech during recording phase (a) while speech is checked for naturalness during confirmation phase (b).

own speech by pressing the REC button. Once the recording is complete, the worker proceeds to the confirmation phase by pressing the NEXT button. During the confirmation phase (Fig. 1(b)), the worker listens to the recording they made and checks whether it sounds natural. If the recording is satisfactory, the worker checks "Are you speaking naturally?" and submits it by pressing the SUBMIT button. If not, the worker can press the BACK button to return to the recording phase.

The worker recorded each sentence three times: once for practice and twice for the real test, with the recording and confirmation phases counted as a single recording. This resulted in a total of 30 recordings as workers recorded for five sentences each in both the lower and upper age group clusters.

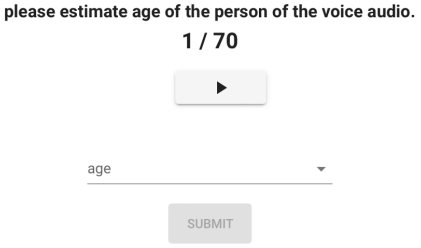


Fig. 2: Interface used by workers to estimate speaker's age includes PLAY button for listening to audio, drop-down menu for selecting age, and SUBMIT button for submitting result.

IV. Age Estimation Experiment

The procedure outlined in Sect. III was employed to record 30 utterances for each of the 50 workers, resulting in a total of 1500 utterances. However, 33 workers did not complete all 30 utterances due to malicious or inadequate recording. In this experiment, those workers were excluded, and age estimation was conducted on the remaining 17 workers' speech using both an automatic age estimation system and manual evaluation.

A. Automatic Age Estimation (AAE)

To conduct age estimation, we utilized an AAE system based on TDNNs, as described in [3]. The TDNNs employed were trained using NIST SRE 2004-2006, Switchboard, Switchboard II Phases 2 and 3, and Switchboard II Cellular Part 1. They were then fine-tuned using the AgeVoxCeleb training dataset.

B. Manual Age Estimation

Workers on MTurk were tasked with estimating the age of a recorded voice, utilizing the UI shown in Fig. 2. Each HIT (the smallest unit of the task requested to workers in MTurk) consisted of 60 voices presented to workers, with 20 voices for qualification testing and 40 for main testing.

In the qualification test, workers were presented with ten utterances from the Fisher corpus, selected to avoid age bias, with five utterances from males and five from females. These utterances were presented twice each to ensure consistency in estimates. We only considered the estimates of workers who provided consistent results for the same utterance and did not respond randomly. Consistency was defined as an error of 10 years or less for the same utterance. The random response was determined by the mean absolute error of the estimated age being 17 years or more because the mean absolute error was 16.25 if workers provided the same age as their answer for all utterances.

During the main test, each worker was presented with a total of 40 utterances that consisted of ten different sentences spoken by four individual workers. These utterances were obtained using the method described in Sect. III. To avoid any order effect, we created three HITs, each containing the same voices but in a different order. Specifically, HIT-1 presented the voices in a random order, with the condition that the same speaker and the same sentences were not presented consecutively. HIT-2 switched the order of the first and second halves of HIT-1, and HIT-3 presented the voices in the reverse order of HIT-1. We expect that this design will reduce any order effects that could potentially impact the results.

V. Effect of Acoustic and Linguistic Information on Age Estimation

We conducted t-tests to determine if there were statistically significant differences in age estimation results *i)* between the two speakers and *ii)* between the two sentences. If we observe a statistically significant difference in estimated ages between the two speakers, we consider it evident that their acoustic characteristics differed enough to enable age discrimination between the two speakers. Conversely, if we observe a statistically significant difference in estimated ages between the two sentences, we consider it evident that the speech content in each sentence had an impact on the estimation results.

A. Age Estimation Results Using AAE System

Figure 3 displays the range of p-values and the corresponding number of speaker pairs falling within each range when the significance test was performed between the estimated ages of two speakers obtained using the AAE system. At a significance level of five percent, a statistically significant difference was observed in 107 out of 136 pairs, indicating that the acoustic features were effective in distinguishing the age of speakers.

Table II lists the p-values resulting from the significance test between the estimated ages of the two sentences. A statistically significant difference was observed in only three out of 45 sentence pairs at a five percent significance level. As the data used in this experiment were chosen to have varying speech content based on the speaker's age, we cannot conclude that the age estimation results produced by the AAE system were significantly affected by the speech content.

B. Manual Age Estimation Results

Figure 4 displays the range of p-values and the corresponding number of speaker pairs. These values were obtained by conducting a significance test between the estimated ages of two speakers obtained using crowdsourced age estimation. When the significance level was set at five percent, statistically significant differences were found in 76 out of 120 pairs. Hence, we can conclude that human age estimation using crowdsourcing is capable of distinguishing the age of 63.3% of speaker pairs based on the acoustic information of the speaker.

Table III lists the p-values for testing significant differences between the estimated ages of the two sentences. When the significance level was set at five percent, no statistically significant differences were found for any pair of sentences. Therefore, it can be concluded that the speech content did not affect the age estimation results when crowdsourcing was used for age estimation.

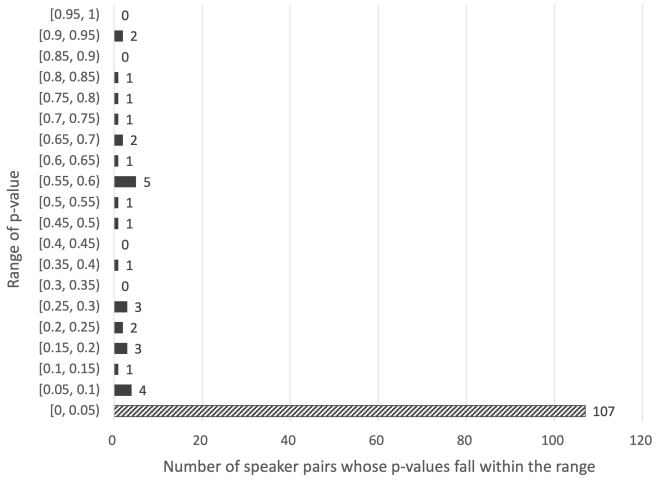


Fig. 3: Effect of acoustic information on automatic age estimation results: Range of p-values and number of speaker pairs falling within each range when testing for significant differences between estimated ages of two speakers. Shaded bar represents number of speaker pairs with statistically significant differences. Among 136 pairs tested, 107 pairs (78.7 %) had statistically significant differences at 5% significance level, indicating that acoustic information was effective in distinguishing between ages of speakers.

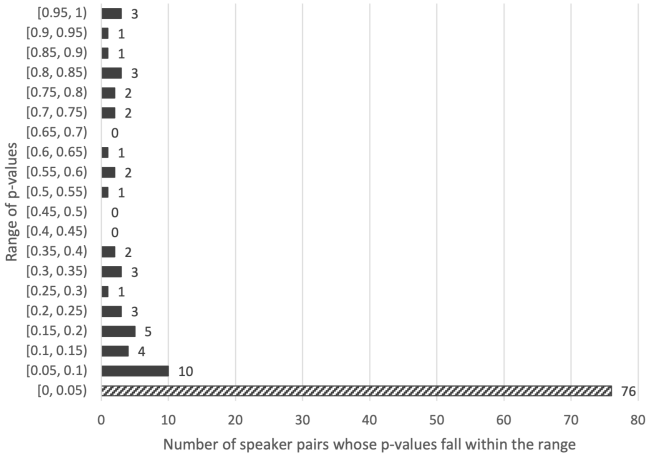


Fig. 4: Effect of acoustic information on manual age estimation results: Range of p-values and number of speaker pairs falling within each range when testing for significant differences between estimated ages of two speakers. Shaded bar represnts number of speaker pairs with statistically significant differences. 76 out of 120 pairs (63.3%) had statistically significant differences at 5% significant level, indicating that acoustic information could be used as basis for age estimation.

TABLE II: Effect of speech content on automatic age estimation results: p-values for significance test between estimated ages of two sentences. Significance level was set at 5%. Underlined numbers indicate significant differences. In all but three cases, no statistically significant differences were found, indicating that content of speech did not significantly affect age estimation results.

					younger generation-biased sentence				older generation-biased sentence				
younger generation-biased sentence	**1**	-											
	2	0.67	-										
	3	0.04	0.06	-									
	4	0.26	0.49	0.27	-								
	5	0.19	0.48	0.15	0.87	-							
older generation-biased sentence	**1**	0.10	0.32	0.37	0.56	0.47	-						
	2	0.26	0.24	0.39	0.70	0.63	0.83	-					
	3	0.30	0.44	0.02	1.00	0.92	0.51	0.72	-				
	4	0.04	0.09	0.74	0.21	0.22	0.55	0.45	0.30	-			
	5	0.05	0.05	0.76	0.17	0.19	0.64	0.44	0.33	0.97	-		
		1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**		

TABLE III: Effect of speech content on results of manual age estimation: p-values for significance test between estimated ages of two sentences. Significance level was set to 5%. No statistically significant differences were found for all pairs of sentences, indicating that age estimation results were not affected by content of speech.

					younger generation-biased sentence				older generation-biased sentence				
younger generation-biased sentence	**1**	-											
	2	0.96	-										
	3	0.91	0.88	-									
	4	0.80	0.78	0.89	-								
	5	0.48	0.46	0.58	0.64	-							
older generation-biased sentence	**1**	0.78	0.82	0.73	0.65	0.34	-						
	2	0.55	0.49	0.61	0.66	0.98	0.31	-					
	3	0.15	0.14	0.22	0.24	0.47	0.09	0.44	-				
	4	0.54	0.59	0.66	0.73	0.90	0.45	0.92	0.39	-			
	5	0.91	0.93	0.84	0.76	0.48	0.90	0.47	0.09	0.53	-		
		1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**		

VI. Conclusion

This study aimed to examine whether the accuracy of age estimation is affected by the content of speech. To achieve this, we used our own dataset of speech with the same content, spoken by speakers of varying ages. Our findings indicated that the speech content did not have a statistically significant effect on age estimation accuracy for either the automatic age estimation system or manual age estimation.

References

[1] S. Si, J. Wang, J. Peng, and J. Xiao, "Towards speaker age estimation with label distribution learning," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4618–4622.

[2] A. I. Mansour and S. S. Abu-Naser, "Classification of age and gender using ResNet-deep learning," *International Journal of Academic Engineering Research*, vol. 6, no. 8, pp. 20–29, 2022.

[3] N. Tawara, A. Ogawa, Y. Kitagishi, and H. Kamiyama, "Age-VOX-Celeb: Multi-modal corpus for facial and speech estimation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6963–6967.

[4] R. Masumura, D. Okamura, N. Makishima, M. Ihori, A. Takashima, T. Tanaka, and S. Orihashi, "Unified autoregressive modeling for joint end-to-end multi-talker overlapped speech recognition and speaker attribute estimation," *arXiv preprint arXiv:2107.01549*, 2021.

[5] S. E. Linville, "Acoustic-perceptual studies of aging voice in women," *Journal of Voice*, vol. 1, no. 1, pp. 44–48, 1987.

[6] R. D. Jacques and M. P. Rastatter, "Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners," *Folia Phoniatrica et Logopaedica*, vol. 42, no. 3, pp. 118–124, 1990.

[7] R. Huntley, H. Hollien, and T. Shipp, "Influences of listener characteristics on perceived age estimations," *Journal of Voice*, vol. 1, no. 1, pp. 49–52, 1987.

[8] A. Braun, "Age estimation by different listener groups," *International Journal of Speech Language and the Law*, vol. 3, no. 1, pp. 65–73, 1996.

[9] L. Cerrato, M. Falcone, and A. Paoloni, "Age estimation of telephonic voices," in *Proceedings of the RLA2C conference (Avignon)*, 1998, pp. 20–24.

[10] M. Brückl and W. Sendlmeier, "Aging female voices: An acoustic and perceptive analysis," in *ISCA tutorial and research workshop on voice Quality: Functions, analysis and synthesis*, 2003.

[11] S. Schötz, *Perception, analysis and synthesis of speaker age*. Lund University, 2006, vol. 47.

[12] N. Tawara, H. Kamiyama, S. Kobashikawa, and A. Ogawa, "Improving speaker-attribute estimation by voting based on speaker cluster information," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6594–6598.

[13] R. M. Hecht, O. Hezroni, A. Manna, R. Aloni-Lavi, G. Dobry, A. Alfandary, and Y. Zigel, "Age verification using a hybrid speech processing approach," in *Proceedings of the INTERSPEECH*, 2009, pp. 184–187.

[14] D. Jiao, V. Watson, S. G.-J. Wong, K. Gnevsheva, and J. S. Nixon, "Age estimation in foreign-accented speech by non-native speakers of english," *Speech Communication*, vol. 106, pp. 118–126, 2019.

[15] S. Saito, Y. Ide, T. Nakano, and T. Ogawa, "VocalTurk: Exploring feasibility of crowdsourced speaker identification." in *Proceedings of the INTERSPEECH*, 2021, pp. 2932–2936.

[16] Y. Ide, S. Saito, T. Nakano, and T. Ogawa, "Can humans correct errors from system? Investigating error tendencies in speaker identification using crowdsourcing," in *Proceedings of the INTERSPEECH*, 2022, pp. 5100–5104.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[18] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.