Perceptual Evaluation of the Effectiveness of Voice Disguise by Age Modification

Rosa González Hautamäki, Anssi Kanervisto, Ville Hautamäki, Tomi Kinnunen

School of Computing
University of Eastern Finland, Finland
{rgonza, anssk, villeh, tkinnu}@cs.uef.fi

Abstract

Voice disguise, purposeful modification of one's speaker identity with the aim of avoiding being identified as oneself, is a low-effort way to fool speaker recognition, whether performed by a human or an automatic speaker verification (ASV) system. We present an evaluation of the effectiveness of *age stereotypes* as a voice disguise strategy, as a follow up to our recent work where 60 native Finnish speakers attempted to sound like an *elderly* and like a *child*. In that study, we presented evidence that both ASV and human observers could easily miss the target speaker but we did not address how *believable* the presented vocal age stereotypes were; this study serves to fill that gap. The interesting cases would be speakers who succeed in being missed by the ASV system, and which a typical listener cannot detect as being a disguise. We carry out a perceptual test to study the quality of the disguised speech samples. The listening test was carried out both locally and with the help of Amazon's Mechanical Turk (MT) crowd-workers. A total of 91 listeners participated in the test and were instructed to estimate both the speaker's *chronological* and *intended* age. The results indicate that age estimations for the intended old and child voices for female speakers were towards the target age groups, while for male speakers, the age estimations corresponded to the direction of the target voice only for elderly voices. In the case of intended child's voice, listeners estimated the age of male speakers to be older than their chronological age for most of the speakers and not the intended target age.

1. Introduction

The human voice is highly flexible [1]. Besides relaying the spoken message, the speaker can alter his or her *voice quality* by changes in phonation, articulation or both [2, 3, 4]. The human voice production system is not rigid and can be modified. Such modifications can be *non-deliberate* or *deliberate* [5]. The former refers to changes in conditions that are not under the speaker's conscious control (*e.g.* speaker's health) or dependent of the environment (*e.g.* Lombard reflex under noisy environments), whereas deliberate modification is actively enforced by the speaker so that he or she is fully aware of it. Voice acting, disguise and impersonation [6, 7, 8] are good examples of this.

Speech modifications are of concern in *speaker recognition* [9], the task of recognizing persons from their voices. With the proliferation of mobile devices, the demand for speech technology applications has increased towards user authentication from a remote terminal. Another application relates to law enforcement and forensics, where a speaker's voice could be used for surveillance or be subjected to forensic voice profiling. Whether performed by an automatic system or a listener, reliability of speaker recognition under deliberate voice modification is of great concern. From the perspective of the perpetrator, the purpose of deliberate voice modification relates to the aim of concealing one's identity. This process, known as *voice disguise*, is the focus of our study.

Speakers may employ a number of strategies to disguise their voices, including the use of external objects (mask, helmet, handkerchief, hand over mouth, pencil, chewing gum); forced modifications of the vocal cavities (pulled cheeks, pinched nose); changes in phonation (creaky, hoar, whisper); adopting a foreign language or dialect. In our recent work, we addressed the impact of voice disguise to automatic speaker verification (ASV) systems [8] and analyzed the variations in acoustic parameters and listeners' performance [10]. The data used in those studies focus on the voice disguise strategy of modifying one's voice to sound like an elderly and a child. Our prior studies found the child mimicry to be particularly detrimental to the performance of ASV systems (see selected results in Table 1) [8], possibly due to a substantial increase in the fundamental frequency that causes a large mismatch in the mel-cepstral features employed by the ASV system. Similar results were found from perceptual experiments, where our listener panel had difficulty connecting modal and disguised samples from same speaker.

Table 1: Performance of i-vector PLDA speaker recognition system, in terms of equal error rate (EER, %), on dataset with male and female speakers speaking in their modal and disguised voices (elderly and child) [10]. Additional results are available in [8].

	Modal	**Disguise elderly**	**Disguise child**
Female	5.05	24.38	31.68
Male	2.82	19.45	30.10

For the corpus collection, the disguise strategy was given to the speakers as an easy-to-understand and easy-to-execute type of speech modification, enabling a common condition to study the vocal variations. However, in our data collection process we did not consider how *believable* the speakers are at producing disguised voices. Therefore, the present study aims to investigate how convincing the disguised samples sound and to evaluate the speaker's ability to disguise their voice by means of age estimations performed by a listener panel.

The perceptual experiment designed for the study aims to answer the following questions:

This work was partially funded by the Academy of Finland (projects no. 309629 and 313970)

- How accurate are human listeners at chronological age estimation from **modal** and **acted** voices?
- How successful are the speakers in modifying their voices in the 'intended' direction, in terms of perceived age by the listener panel?

The goal to evaluate the effectiveness of voice disguise is two-fold. On the one hand, it can give a perspective of how high is the threat of disguised speech to speaker recognition systems and how likely speakers are able to evade recognition with malicious intentions. On the other hand, if the user needs to hide his identity for legitimate reasons, effective disguise could help in protecting the speaker's privacy and identity.

2. Perceptual age estimation from speech

Humans infer speaker characteristics from the voice on daily non face-to-face interactions. For instance, listening to speech relayed through a public address system, radio program, or speech interface gives the listener an impression of the speaker's gender, age, language, dialect, level of education, personality, among other factors. We focus on listener's ability to predict speaker's age from voice.

Age prediction from one's voice has been addressed in a number of previous studies. Goy *et.al.* [11] studied age-related differences between older and younger speakers, and also listeners in terms of perceived speech and voice quality. In their experiments, the listeners estimated age and gender of young and old speakers, along with naturalness, clarity and intelligibility. The perception of voice quality was found to be significantly influenced by the age of the listeners. Vowel samples were used for the age estimation experiments and it was found that their younger listeners were more accurate.

Pettorino and Giannini [12] addressed the degree to which listeners are able to effectively estimate the speakers' age. One of their experiments found that estimating the speaker's age in an unconstrained manner is a difficult task, while classifying a voice directly by age groups was relatively easy.

Age estimation from speech is of interest not only in defining high-level speaker characteristics but in the understanding of the changes related to aging. The aging process is not uniform and several extrinsic factors may affect speaker's voice as he or she ages. The work by Schötz [13] presents an acoustic-phonetic study of the speaker's age. The study examined acoustic parameters of the voice such as speech rate, sound pressure level (SPL) and fundamental frequency ($F0$). These have been found important as acoustic correlates of the speaker's age. However, there is no clear relation between perceptual cues and listeners strategies used in age estimation and the age-related acoustic correlates. Also, other factors related to the speech sample, listening condition and listener's age have been found to have an effect on the human perception of age [13, 14]. In age estimation by listening, it has been found that the age of young speakers tends to be overestimated, while the age of older speakers tend to be underestimated [14, 15].

Previous work has also addressed the effect of age disguise in the estimation of the speaker's age [14, 16]. Skoog and Eriksson [14] studied the voice disguise of speakers that attempt to sound 20 years older and 20 years younger. It was found that the listeners' perceived an age change of 3 years, rather than the expected 20 years. In contrast, our study aims to evaluate whether the voice disguise attempts are perceived in the direction of the target age group, i.e. if the speakers were successful in the intended age modifications.

3. Perceptual experiment

We designed a perceptual test to study the quality of disguised voices using human listeners. The goal was to evaluate the disguise attempts through age estimation based on the speaker's voice. To this end, we first need three different definitions of *age*:

Chronological age: Objective age defined as the person's age at the time of the speech recording. We define this in years.

Perceived chronological age: Subjective age rating by one listener concerning a given speech segment that reflects the listener's best guess of the actual chronological age. Differently from the actual chronological age, we define this age in terms of age categories.

Perceived intended age: Categorical subjective age rating similar to the previous, except for one key difference: it is the listener's best guess of what age the speaker has *intended* to sound like. Such variable can be meaningfully defined only for listeners who are aware of the presence of voice acting.

The listeners chose their estimations of the speaker's perceived chronological age and perceived intended age from five pre-defined age intervals: *child* (younger than 18 years old), *young adult* (approx. 18-30 years old), *middle-age* (approx. 31-64 years old), *retired* (approx. 65-80 years old) and *senior citizen* (older than 81 years old). These intervals were determined empirically to have a balanced division of the speakers' chronological ages. The two boundary choices, younger than 18 years (child) and older than 80 years (elderly), are included so that for any speaker in our data, the listener has a chance to make a 'correct' age estimation in the case of *successful* age modification. For example, for a younger speaker that is able to modify the voice to sound younger, a listener could assign the child category.

Figure 1 shows the distribution of the speakers' chronological ages and Table 2 shows the speakers' distribution in the pre-defined age intervals. The speech data for the listening test was recorded with a close-talking microphone and from the second recording session, where the speakers were generally more comfortable with the tasks. The data is the same as in our recent prior work [8, 10]; more details of the data collection can be found therein.

Table 2: Speakers' age distribution according to the age categories used for the perceptual test.

	Female	**Male**
Younger than 18 years	0	0
18 - 30 years	19	14
31 - 64 years	11	14
65 - 80 years	1	1
Older than 81 years	0	0

A set of 540 speech segments were selected for the perceptual test corresponding to equal number of segments from 60 native Finnish speakers in their modal, elderly and child voices. Three utterances were selected from each voice type per speaker. Therefore, 60 speakers $\times$ 3 voice types $\times$ 3 utterances = 540 utterances in total. The speech segments from all

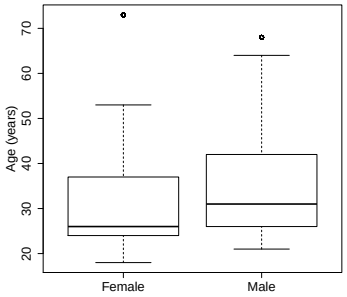


Figure 1: *Speakers' chronological age distribution per gender (31 female and 29 male).*

60 are distributed in 12 speaker-disjoint folds, each consisting of 45 speech segments, from different speakers, with the same number of segments in modal and disguised voices (15 utterances $\times$ 3 voice types = 45). The 12 folds are then presented, likewise, to 12 independent listener panels. The listeners were informed that *all* the speech segments contain voice acting — in this way, the listeners will focus on the perceived age estimation regardless of whether a particular sample is acted or not.

All the listener responses were collected with the aid of crowdsourcing. We had two groups of crowdworkers: those recruited ourselves, and those recruited through a crowdsourcing service. Concerning the recruited listeners by the authors, we prepared an online survey form that was completed by a total of 22 listeners. Each participant assessed 45 speech samples. Even though such listeners participation is not considered a laboratory test, the listeners have participated in our previous voice comparisons tests, hence we regard them as reliable collaborators.

The advantage of using a paid crowdsourcing service is to reach a larger number and more diverse pool of participants in a short time. To this end, 69 listeners participated in the listening test via Amazon mechanical turk[1] (AMT) service. From these listeners, 35 listened to all the 45 samples (the same number as the recruited listeners), 23 listeners assessed more than 10 samples, and the remaining listeners less than 8 samples. All the listeners from the AMT group are non-native Finnish speakers while 11 listeners from the recruited group are native Finnish speakers.

4. Results

Listeners estimated speaker's age-group (five in total) for each speech sample. However, we are more interested in the perceived age in years rather than the number of votes per age group. We resort to estimate the *expected perceived age* per speaker, using the number of votes per age group as a weight. Let x_i be the center of mass of age group $i \in \{1, 2, \ldots, N\}$, v_i number of votes in age group i and V the number of all votes for this speaker. In our case, the total number of age groups is $N = 5$. The expected perceived age a for a single speaker is then defined as:

$$a = \frac{1}{V} \sum_{i=1}^{N} x_i v_i. \quad (1)$$

[1] https://www.mturk.com/

Rest of this study will consider this as the age estimate given by the listeners collectively.

4.1. Listener accuracy

We evaluated the estimations obtained by the listeners recruited by the authors (UEF) and the ones from crowdsourcing (AMT). This comparison provides information of the similarity of the responses by both groups of listeners. Using the modal voice samples, the mean perceived chronological age for each speaker was obtained using Eq. (1), then the correlation between the two listeners groups estimations was calculated. Figure 2 was generated using `ggpubr` R package. It presents the correlation using *Pearson* method where UEF listeners and AMT listeners estimations are significantly correlated with a coefficient of 0.72 and *p-value* of $1.2e^{-10}$. Indicating a positive correlation between the variables. The gray area shows the uncertainty of the correlation coefficient at the 95% confidence interval.

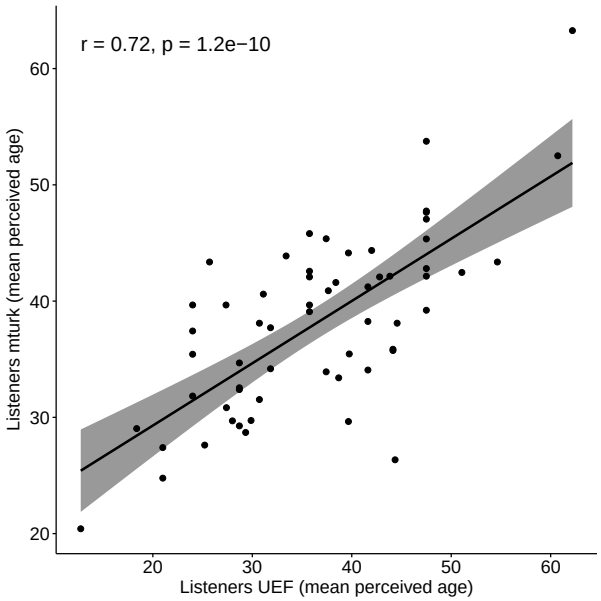


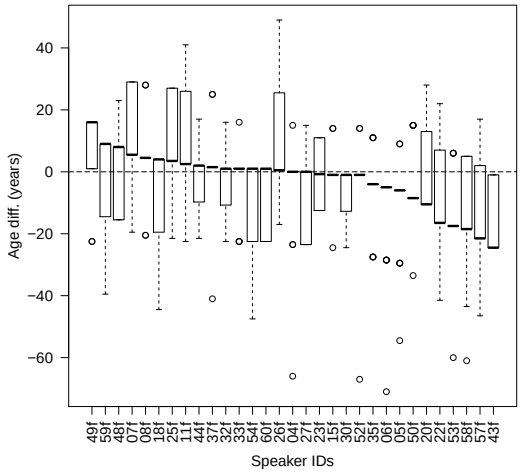
Figure 2: *Pearson correlation of the speakers' perceived chronological age for the modal voice by the recruited listeners (UEF) and listeners from the crowd-source platform (Amazon mechanical turk). Correlation coefficient* 0.72*, p-value* $1.2e^{-10}$ *with a confidence interval of* $[0.5658\ ,\ 0.8212]$ *at 95%.*

Even though the listener groups are of different sizes, the perceived age estimations themselves are comparable. We can therefore expect similar performance between random listeners from the UEF and the AMT groups. For the rest of the analysis, therefore, we pool the recruited and the AMT listeners into one listener panel of 91 listeners.

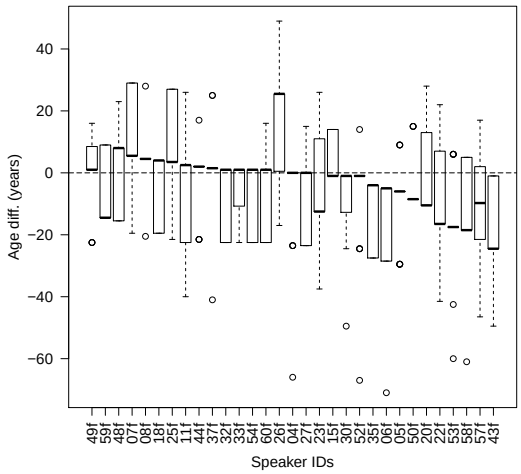
To evaluate the age estimations for each speech sample, the *age difference* was calculated as follows:

$$\textit{Age diff.} = \textit{Chronological age} - \textit{Perceived age},$$

where *perceived age* corresponded to either the perceived chronological age or to the perceived intended age. A positive age difference can be interpreted as the perceived age being underestimated or lower than the speaker's chronological age,

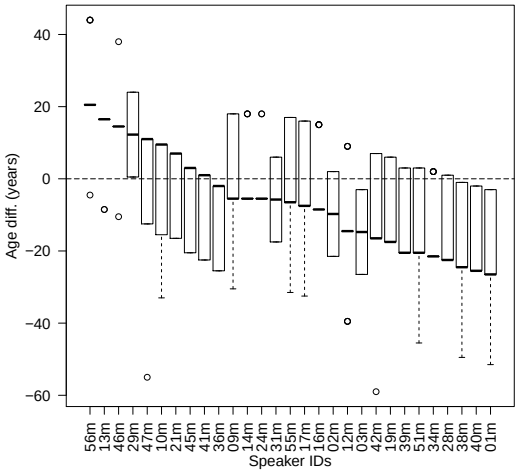


(a) Difference for chronological age estimates from modal voice

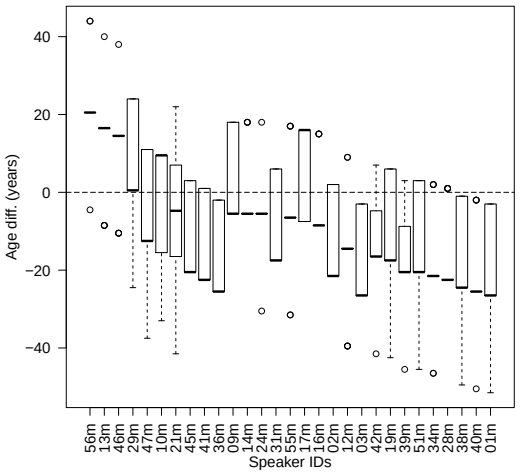


(b) Difference for perceived intended age from modal voice

Figure 3: Female speakers age difference in years between the speakers' chronological age and the perceived chronological and perceived intended age estimates for the modal voice segments. The speakers are ordered by the median age difference in descending order. The graphs show small differences indicating that the estimates for the age of the speaker and the "intended" age are close for modal voice.



(a) Difference for chronological age estimates from modal voice



(b) Difference for perceived intended age from modal voice

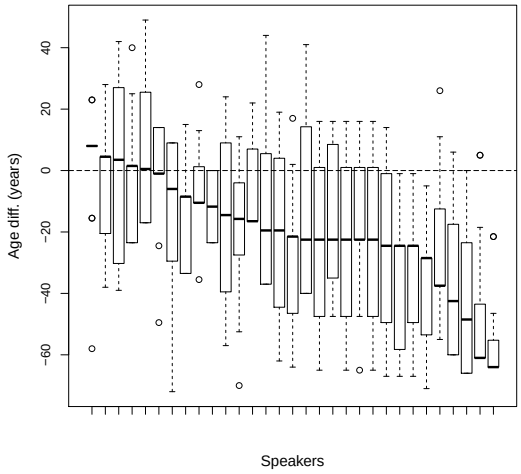
Figure 4: Male speakers age difference in years between the speakers' chronological age and the perceived chronological and perceived intended age estimates for the modal voice segments. The speakers are ordered by the median age difference in descending order. The graphs show a similar pattern in the age differences, in both graphs the estimations are mostly to the negative side indicating that the age of the speakers is overestimated.

and a negative value would correspond to an overestimated perceived age, or as higher than the speaker's age.

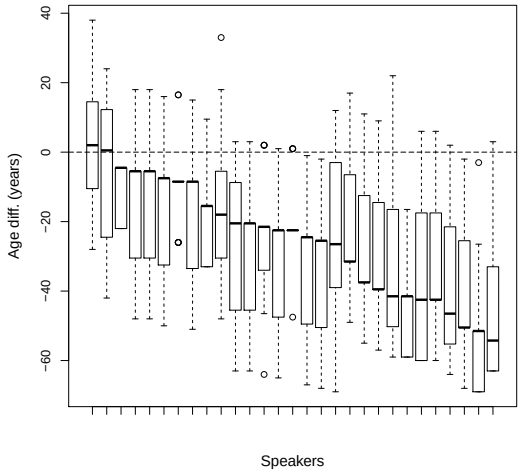
To evaluate listeners ability to estimate speaker's chronological age, we compared estimated chronological and intended ages to the real age of the speaker only for the modal voice speech segments. The assumption is that for each speaker the perceived chronological age and the perceived intended age will be similar or show a small difference, as these samples do not include disguised voices.

Figure 3 shows the perceived age estimations for female speakers in the case of modal voices. The median age differences are close to the zero difference region for many speakers. This result indicates that the listeners' perceived age estimations are close to the speakers' chronological ages and that the perceived chronological and intended age estimations are similar for the modal voices. This agrees with our assumption for most of the speakers.

The results for male speakers are shown in Figure 4 where the differences for perceived chronological (Fig. 4 (a)) and intended age (Fig. 4 (b)) follow the same pattern but, in contrast to female speakers, the difference is not close to the zero differ-

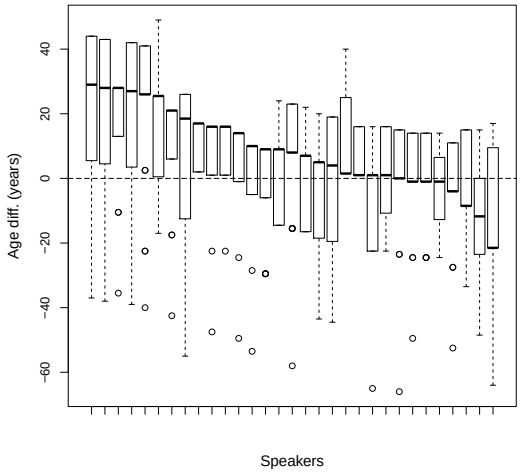


(a) Female

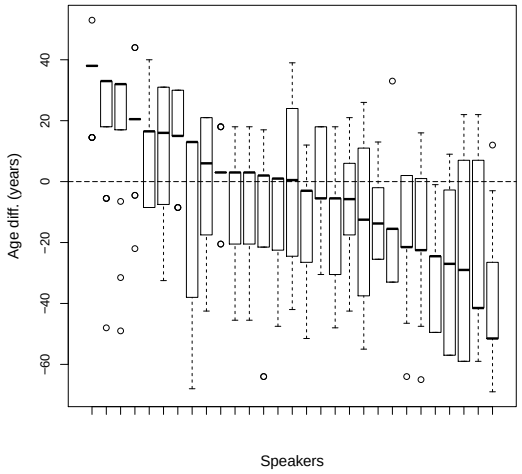


(b) Male

Figure 5: Age difference between the speakers' chronological age and perceived intended age for the intended *elderly* voice segments per gender. The speakers are ordered by the median age difference in a descending order.



(a) Female



(b) Male

Figure 6: Age difference between the speakers' chronological age and perceived intended age for the intended *child* voice segments. The speakers are ordered by the median age difference in a descending order.

ence region and it is negative for most of the speakers. This indicates that for most of the speakers, the listeners over-estimated the ages, which is a common problem on age estimation of young speakers.

Age estimation from speech is a difficult task and our experiment has the additional challenge of including voice disguise by age modification. For our corpus, our listeners were better at estimating the chronological age of female speakers than male speakers. Even though the age estimation show variations for each speaker, we can consider the performance of the listeners to be consistent in the estimation of chronological age from modal and disguised voices for most of the speakers.

4.2. Perceived age of disguised voices

Similar to Fig. 3 (b), let us now consider the samples containing disguised voices. The assessments were carried out using the age difference between the chronological age of the speakers and the perceived intended age. In this way, the listeners' age estimations will reveal whether the speakers' attempts reached the intended voice.

Figure 5 shows the age difference in years for the speech samples with intended elderly voice. For female speakers (Fig. 5 (a)), the mean age difference range is from 2.46 to -57.12 with a mean of -19.45, while for male speakers (Fig. 5 (b)), the range is from 1.36 to 46.53 with a mean of -24.97. We observe a negative difference for both genders, suggesting that

the listeners estimated the perceived intended age to be older than the speaker's chronological age. This can be considered as a successful age disguise attempt in terms of the perceived age. Even though most of the speakers were able to sound older than themselves, the age difference is small to reach the intended age (elderly or older than 80 years old).

For the intended child voice, the age difference is expected to be positive to be considered successful age disguise: this means that the age of the perceived intended voice is younger than the speaker's chronological age. Figure 6 shows the age difference per gender for intended child voice. For female speakers (Fig. 6 (a)), the mean age difference range is from 20.81 to -11.29 with a mean of -5.08, while for male speakers (Fig. 6 (b)), the range is from 33.61 to -38.93 with a mean of -5.95. For female speakers, most of the perceived intended age estimates are below the speaker's chronological age, but just a few were in an age difference that would correspond to the child's voice. This was not the case for male speakers, where the perceived intended age was overestimated with respect to the speaker's chronological age. In other words the listeners' estimations indicate that the voices corresponded to older voices and not the intended child voices.

4.3. Chronological age estimation from disguised voices

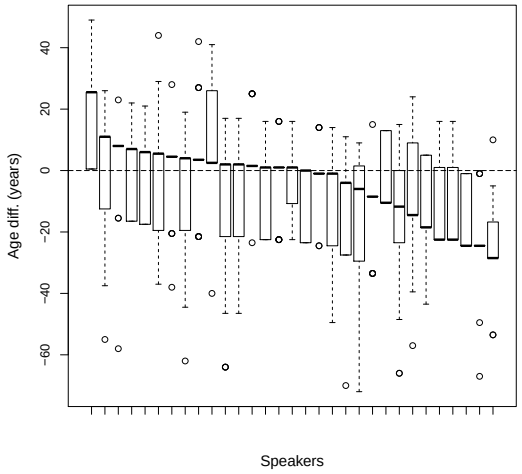
Another interesting point was to evaluate the estimation of the speaker's chronological age when he or she disguises the voice. This could give insights of how challenging the age estimation task is when the speaker is trying to disguise the voice by means of age modification. Our results, Figures 7 and 8, show that for most of the male and female speakers, the perceived chronological age estimations were similar for both attempts of elderly and child voices. The results indicate that the listeners' age estimation only were affected for a few speakers but followed the results obtained from the perceived chronological age from modal voices.

5. Conclusions

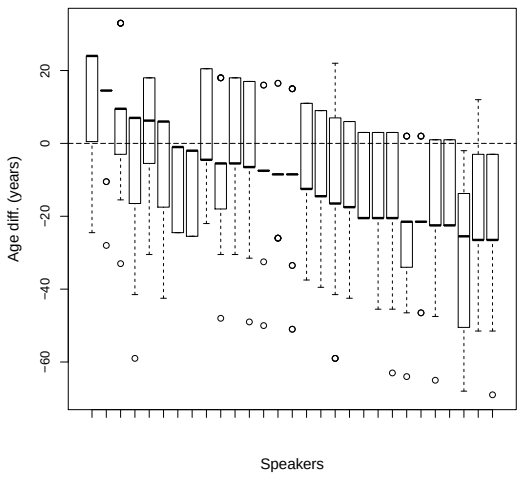
In this study, we looked into what a successful voice disguise attack on automatic speaker verification sounds like to humans. We approached this question by gathering human age estimations for modal and disguised speech samples and comparing these estimated ages against each other. Listeners estimated the chronological and the intended age of the speaker for the selected speakers' samples. This study also served as the authors' first perceptual experiment with a crowdsourcing platform, Amazon Mechanical Turk, to facilitate the convenient collection of listener results. We observed a positive correlation between the AMT responses to those collected locally with ad-hoc listener recruitment. This gives us confidence to the use of paid crowdsourcing services for future studies as well.

In our experiments, listeners were able to approximate the chronological age of female speakers with their modal voices, while for male speakers, the age was systematically overestimated. This is a common observation in previous studies for the age estimation of young speakers [15], which is the case of most of the speakers in this study. Also, the perceived chronological age from disguised voices followed similar estimations as with modal voice.

In the evaluation of the disguised voices for female speakers, the listeners' estimations of the perceived intended age followed the direction of the target age for intended elderly and child voices. That was also the case of intended old voice for male speakers. However, the intended child voice was perceived as belonging to an old person for most of the male speakers. This indicates that even though intended child voice from most male speakers did not sound believable, this type of disguise affected the performance of ASV systems.



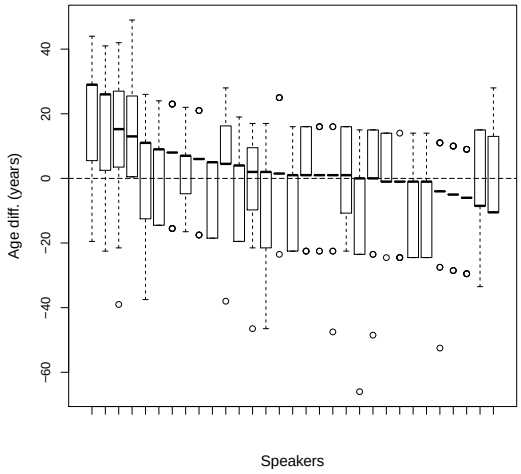
(a) Female



(b) Male

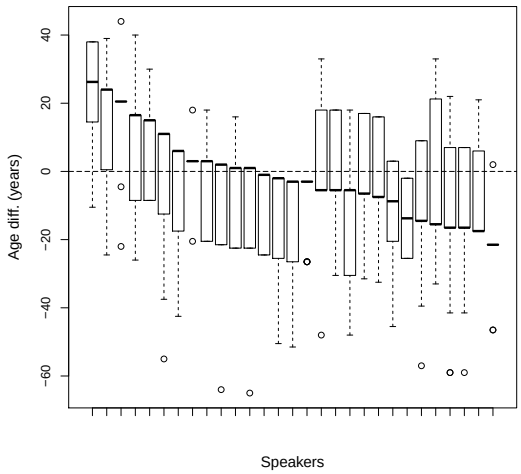
Figure 7: Age difference between the speakers' chronological age and perceived chronological age for the intended *elderly* voice segments. The speakers are ordered by the median age difference in a descending order.

In the light of these results, future work could focus on the cues humans detect from disguised speech to effectively identify that it is disguised. In general, humans are able to distinguish between acted and non-acted voice, but a standard automatic speaker recognition system is not equipped to recognize speakers under these type of disguise strategies.



(a) Female

."



(b) Male

Figure 8: Age difference between the speakers' chronological age and perceived chronological age for the intended *child* voice segments. The speakers are ordered by the median age difference in a descending order.

6. References

[1] Anders Eriksson and Pär Wretling, "How flexible is the human voice – a case study of mimicry," in *Proc. European Conf. on Speech Commun. and Technol. (EUROSPEECH)*, Rhodes, Greece, 1997, vol. 2, pp. 1043–1046.

[2] Anders Eriksson, "The disguised voice: imitating accents or speech styles and impersonating individuals," vol. 8, pp. 86–96. Edinburgh University Press, 2010.

[3] Eugenia San Segundo, Helena Alves, and Marianela Fernández Trinidad, "CIVIL corpus: Voice quality for speaker forensic comparison," *Procedia-Social and Behavioral Sciences*, vol. 95, pp. 587–593, 2013.

[4] Talal Bin Amin, Pina Marziliano, and James S. German, "Nine voices, one artist: Linguistic and acoustic analysis," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, July 2012, pp. 450–454.

[5] Robert Rodman and Michael Powell, "Computer recognition of speakers who disguise their voice," in *The international conference on signal processing applications and technology (ICSPAT2000)*, 2000.

[6] Adrian Leemann and Marie-Jos Kolly, "Speaker-invariant suprasegmental temporal features in normal and disguised speech," *Speech Communication*, vol. 75, pp. 97–122, 2015.

[7] Cuiling Zhang, "Acoustic analysis of disguised voices with raised and lowered pitch," in *Proc. Int. Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2012, pp. 353–357.

[8] Rosa González Hautamäki, Md Sahidullah, Tomi Kinnunen, and Ville Hautamäki, "Age–related voice disguise and its impact in speaker verification accuracy," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2016, pp. 277–282.

[9] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[10] Rosa González Hautamäki, Md Sahidullah, Ville Hautamäki, and Tomi Kinnunen, "Acoustical and perceptual study of voice disguise by age modification in speaker verification," *Speech Communication*, vol. 95, pp. 1–15, 2017.

[11] Huiwen Goy, M. Kathleen Pichora-Fuller, and Pascal van Lieshout, "Effects of age on speech and voice quality ratings," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 1648–1659, 2016.

[12] Massimo Pettorino and Antonella Giannini, "The speakers age: A perceptual study," *Proceedings of ICPhS XVII, Hong Kong*, pp. 1582–1585, 2011.

[13] Susanne Schötz, *Acoustic Analysis of Adult Speaker Age*, pp. 88–107, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[14] Sara Skoog Waller and Mårten Eriksson, "Vocal age disguise: The role of fundamental frequency and speech rate and its perceived effects," *Frontiers in Psychology*, vol. 7, no. 1814, 2016.

[15] Sara Skoog Waller, Mårten Eriksson, and Patrik Sörqvist, "Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age," *Frontiers in psychology*, vol. 6, 2015.

[16] Norman J. Lass, Leah A. Justice, Brenda D. George, Linda M. Baldwin, Kathleen A. Scherbick, and Deborah L. Wright, "Effect of vocal disguise on estimations of speakers' ages," *Perceptual and Motor Skills*, vol. 54, no. 3_suppl, pp. 1311–1315, 1982.