TOWARDS CONTROLLING FALSE ALARM — MISS TRADE-OFF IN PERCEPTUAL SPEAKER COMPARISON VIA NON-NEUTRAL LISTENING TASK FRAMING

Rosa González Hautamäki and Tomi H. Kinnunen

Computational Speech Group, School of Computing, University of Eastern Finland
P-O-BOX 111, 80110 Joensuu, Finland
rgonza@cs.uef.fi, tkinnu@cs.uef.fi

ABSTRACT

Speaker comparison by listening is a valuable resource, for instance, in human voice discrimination studies, and voice conversion (VC) systems evaluations. Usually, listeners are provided with application-neutral guidelines that encourage retaining overall high speaker discrimination accuracy. Nonetheless, listeners are subject to misses (declaring same-speaker trial as different-speaker) and false alarms (vice versa) with possibly non-symmetric outcomes. In automatic speaker verification (ASV) applications, the consequences of a miss and a false alarm are rarely equal, and decision making policy is adjusted towards a given application with a desired miss/false alarm trade-off.

We study whether listener decisions could similarly be controlled to provoke more accept (or reject) decisions, by framing the voice comparison task in different ways. Our *neutral*, *forensic*, *user-convenient bank* and *secure bank* scenarios are played by disjoint panels (through Amazon's Mechanical Turk), all judging the same speaker trials originated from RedDots and 2018 Voice Conversion Challenge (VCC 2018) data. Our results indicate that listener decisions can be influenced by modifying the task framing. As a subjective task, the challenge is how to drive the panel decisions to the desired direction (to reduce miss or false alarm rate). Our preliminary results suggest potential for novel, application-directed speaker discrimination designs.

Index Terms— Speaker verification, speaker discrimination, listener performance, decision making

1. INTRODUCTION

In our everyday life, we recognize people from their voices seamlessly. In specific, humans are good at recognizing *familiar* (known) voices [1], such as family members and friends. This ability is suspected to play a role in evolution as one of the survival mechanisms to differentiate trusted parties from potential enemies. Infants prefer the voice of their mother over strangers [2]. In contrast to known speaker recognition, *unknown* speaker recognition involves differentiating two voices from each other. Since [3], it is acknowledged that the neurophysiological mechanisms underlying familiar and unfamiliar speaker recognition are generally different. We point the interested reader to [4, 5] for recent reviews. We focus on unfamiliar speaker recognition.

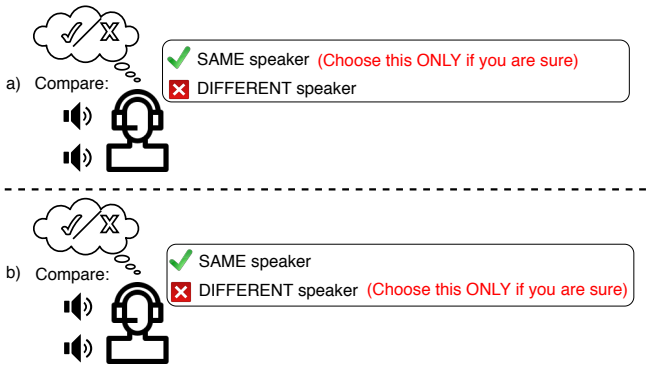

Fig. 1. An example of non-neutral framing of a voice comparison task, with the goal to purposefully influence the listener decisions. The upper task framing (a) is designed to hint the listener that false alarm (false acceptance) incurs high penalty, while the lower task framing (b) serves the opposite purpose.

The task of comparing two voices to decide whether or not they were spoken by same or different individuals (regardless of *who* those individuals are) goes by different names depending on the field; in cognition studies, it is known as *speaker discrimination*, in engineering as *speaker verification* or *speaker detection*, and in forensics as *forensic voice comparison*. Besides its relevance to cognition studies, forensics, and integration of human decisions with *automatic speaker verification* (ASV) [6, 7, 8, 9], speaker discrimination is used in evaluating text-to-speech (TTS) and voice conversion (VC) systems in their ability to mimic voice characteristics of targeted speakers [10]. The usual setting involves gathering speaker similarity (and quality) ratings from many listeners and presenting the cumulative or averaged results. Given the labor associated with data collection in controlled laboratory environments, the focus has shifted towards the use of

This research was partially funded by the Academy of Finland, projects no. 309629. The authors thank V. Hautamäki from the statistical consulting services at the University of Eastern Finland.

crowdsourcing.

Depending on the application, the decision in a speaker discrimination task can be *soft* or *hard*. While a hard *same/different* decision is necessary in many applications, soft decision enables reporting one's uncertainty of the decision. Listeners are typically asked to report a verbal, ordinary scale value. The verbal scale, guidelines and expected standards for the listening setup differ. Untrained, or *naive* listeners perform speaker discrimination on intuitive ground and are typically used in TTS and VC studies. This is in contrast to *expert* listeners who are trained for highly specialized tasks such as forensic voice comparison, and who combine their domain expertise with automatic and statistical methods [11]. We focus on naive listeners.

In contrast to neurophysiological studies of speaker discrimination [4, 5], we focus on high-level cognitive *decision making*. Even if not typically framed as such, speaker discrimination is a decision problem in the face of uncertainty and risks associated with the decisions (actions). For instance, false acceptance (*false alarm*) would be costly in forensics, while false rejection (*miss*) of genuine user in a customer service application would annoy the user. While ASV researchers are accustomed to think trade-offs between misses and false alarms, this perspective seems to have received limited attention within auditory speaker discrimination. This study serves to bridge the gap and, as we hope, to demonstrate a possibly missed dimension in speaker discrimination task designs.

Our proposal, in brief, is to provide listening instructions that encourage the subjects to move their role from a dull voice comparator towards a responsible decision maker in an envisioned application, framed as a role play. While this is not the first study to use role play (*e.g.*, [12, Section V.E]), our novelty is that of intentional biasing of listener decisions along a hypothetical risk space. We hypothesize that, if the applications are framed to invoke attitudes (positive or negative) or mental pictures, listeners could be driven to alter their false alarm and/or miss rates (SEE FIGURE 1). Should this be shown to be plausible, one might benefit from it so as to drive VC or TTS tests towards a specific application. We use Amazon Mechanical Turk crowdworkers who rate trials from two widely adopted dataset, RedDots [13] and 2018 Voice Conversion Challenge (VCC 2018) [10]. The former represents text-dependent speaker verification in a mobile context and the latter represents high-quality controlled audio — with inclusion of advanced voice conversion 'attack' (system N10) with the aim to mislead the uninformed listener.

2. SPEAKER COMPARISON AS A DECISION TASK

Even if our focus is on human decisions, we provide a brief review of decision making principles in automatic speaker verification (ASV) systems to serve as inspiration.

2.1. Automatic speaker verification systems

An ASV system is a hypothesis testing machine that accepts a pair of speech utterances (one with *known* speaker identity, collected at user enrolment or training stage; another one with *unknown* identity, collected at the verification stage), say $x = (x_{\text{enroll}}, x_{\text{test}})$, and outputs a numerical *detection score* $s \in \mathbb{R}$ using a model described by parameters $\vec{\theta}_{\text{asv}}$. While the unit/scale of s is arbitrary, high scores indicate stronger support for *same speaker* (target) hypothesis and low scores for the alternative (nontarget) hypothesis. The score, often *logarithmic likelihood ratio* (LLR) from a statistical model (*e.g.*, [14]), is a *soft* decision. It is converted to hard decision by choosing an *action* $\alpha \in \mathcal{A} = \{\texttt{ACCEPT}, \texttt{REJECT}\}$, in practice, by comparing s to a pre-determined threshold t. The speaker is accepted if $s > t$, and rejected otherwise.

The choice of t impacts the trade-off of the two possible types of detection errors, *false alarms* (choosing ACCEPT when the actual speaker identities differ) and *misses* (choosing REJECT when the speaker identities are the same).The false alarm rate $P_{\text{fa}}(t)$ and the miss rate $P_{\text{miss}}(t)$ are, respectively, decreasing and increasing functions of t. If one needs a high-security application (low P_{fa} desired), it comes with a trade off in increased miss rate (high P_{miss}), a measure of user inconvenience — and vice versa. In practice, the threshold is set to balance the two error rates with the aid of a cost model that takes into account the losses incurred by making wrong decisions, as well as the *a priori* expected occurrence of target speaker. The *detection cost function* (DCF) endorsed by the *National Institute of Standards and Technology* (NIST) in their speaker recognition evaluation campaigns [15] formalizes this notion:

$$\text{DCF}(t) = C_{\text{miss}}\pi_{\text{tar}}P_{\text{miss}}(t) + C_{\text{fa}}(1-\pi_{\text{tar}})P_{\text{fa}}(t),$$

where $C_{\text{miss}} > 0$ and $C_{\text{fa}} > 0$ are, respectively, costs incurred by making a miss and a false alarm, and π_{tar} is prior probability of target user. These costs are selected differently depending on the application.

2.2. Humans

To summarize the above discussion, in ASV applications, one is rarely interested in the overall accuracy, but prefers either low false alarm rate or low miss rate region. This raises a question whether similar ideas could be applied in the case of listeners. That is, can we purposefully influence decisions of the listener so as to increase (or decrease) their miss or false alarm rate? If so, does that come with a trade-off in the other error rate, compatible with the idea of some hypothesized monotonically related error trade-off curve? We address these questions in the form of a *voice comparison role play*. Our work is in part inspired by [12, Section V.E], where listeners, judging speaker similarity of various state-of-the-art TTS and VC systems at the time, were asked to imagine being responsible to grant or deny access to bank accounts:

> "*listeners were ... informed that they would only have a short recording of a person's voice to base their judgement on. It was stressed that it was important to not give access to 'impostors' but equally important that access was given to the 'bank account holder'*.

By saying 'equally important' the authors might have implied that the listeners should aim at minimizing their total error rate, or a DCF-like cost function with no strong preferences on either detection error: $C_{\text{miss}} \approx C_{\text{fa}}$ and $\pi_{\text{tar}} \approx 0.5$. One may argue such instructions to present mapping from verbal instructions to some hypothetical cost function (even if no exact numerical values can be asserted) that a listener should follow. We make an attempt to sample listener decisions from not only the 'neutral' decision region, but low miss rate and low false alarm rate regions, to avoid making decisions that are perceived to have more costly consequences.

It is important to remark that human decision making does not follow the same objective rules as decision making in ASV systems. The DCF framework of NIST stems from Bayes' minimum risk classification [16, 17] and encourages the decision maker to make a choice that minimizes *expected* loss (or maximizes expected *utility*). Unfortunately, this is *not* how humans make choices. Within psychology and economics, it is widely established that human decision making involving probabilities and risks does not follow maximization of expected utility. For instance, the widely-celebrated *cumulative prospect theory* [18], originating initially [19] from a series of gambling experiments, demonstrates a number of risk patterns that do not follow linear expectations, rather, non-linearly transformed probabilities, with *subjective value function* that quantifies subjective value of *losses* and *gains* relative to some reference. Generally, losses and gains are perceived differently by humans. All our role plays are framed in terms of losses, rather than gains.

3. VOICE COMPARISON ROLE PLAY SCENARIOS

Given a speaker discrimination task scenario, listeners are asked to decide whether two audio samples correspond to same or different speakers. The aim is to evaluate different listeners panels' decision process guided by different instructions (framing). We consider four role play scenarios detailed as follows. The *neutral* scenario serves as a reference case. The aim of the *forensic* and *secure bank* scenarios is to provoke listeners to reduce false alarms, while *user-convenient bank* serves to reduce miss rate. While disjoint listening panels play each role, *all of them process the same trials*. Thus, assuming the listening panels behave similarly, systematic change in the observed miss and/or false alarm rates between groups would indicate that the instructions influence the decisions. Our voice comparison role plays are intended to represent a sampling of typical speaker recognition applications. We stress that our interest is the impact of *framing* to listener decisions, with the aid of imagined applications. Such procedures should not be used in real-world high-stakes applications. To take ethical consideration into account, we provided the following disclaimer: *We emphasize that you play a role game: none of the samples that you will listen represent (real crime cases / customers from a banking domain), and none of your suggestions are going to be used for any legislative purposes; all the samples that you will hear are from voluntary participants.*

3.1. Research hypotheses

Our research hypotheses concern pairwise group differences:

- **H1.** Forensic group has lower P_{fa} than neutral group.
- **H2.** User-convenient bank group has lower P_{miss} than neutral group.
- **H3.** Secure bank group has lower P_{fa} than neutral group.
- **H4.** Secure bank has lowered P_{fa} *and* increased P_{miss} relative to user-convenient bank.

The first three are anticipated based on framing of the role play. The last one, in turn, derives from the analogy from ASV systems where miss and false alarm rates are traded off.

3.2. Framing setup 1

The relevant sentences concerning listener biasing are highlighted. We hope that the reader finds these instructions intuitive.

NEUTRAL. *Your task is to decide whether or not the same speaker is present in the two recordings. Please do your listening as carefully as possible.*

FORENSIC. *Your task is a role play game that involves listening to pairs of speech samples. Your task is to decide, as accurately as possible, whether or not the same speaker is present in the two recordings. Put yourself in the shoes of a* ***crime investigator*** *at a police station. One of the samples represents a criminal's voice at a crime scene (e.g., fraudster call or a robbery recorded by a surveillance camera). The other one is suspect's voice recording at a police station. Remember the phrase "innocent until proven guilty" – wrong judgement could lead to an innocent person being convicted. Therefore, please do your listening as carefully as you can.*

USER-CONVENIENT BANK. [Omitted common part] *Put yourself in the shoes of an* ***intelligent voice banking service*** *that handles a large number of transactions daily. The customers use smart-phones to interact with the service using voice only* (e.g., *"please pay Bob $150", "I approve this payment" or "What is my saving account's balance?"). The services use* ***voice authentication technology*** *to first verify the caller's identity based on a pass-phrase. One of the samples you listen is the customer's earlier verified call and the other one is a new call. In no circumstances, you should reject the real account owner, as this could*

lead to customer inconvenience (and the bank to lose the customer) [Omitted common part].

SECURE BANK. [Omitted common part] *Please be alerted that the bank service might also be accessed by unauthorized hackers. In no circumstances, you should accept a wrong person, as this could lead to money loss (and serious loss of bank's reputation).* [Omitted common part]

3.3. Framing setup 2

Taking into consideration crowdsourcing's framework, where more *workers* (listeners) can be reached to execute several tasks in a relative short time, a key aspect in describing the task is brevity. Another setup, to test the task framing influence in the decisions, consists of a simplified version of the task in the form of bullet points, images and added warning notes to specify the errors the role play tries to minimize. The scenarios are the same as in framing setup 1. As an example, the forensic scenario instructions that fit the framing decision in Figure 1a) are as follows:

FORENSIC. In this task, you will be presented with an audio pair:

- Listen carefully to both recordings
- Imagine your answers are used to surveil criminal activity.

Your decision:

SAME person (Select this ONLY if you are sure)
DIFFERENT person

Similar decision framing is used in Secure bank where the idea is to minimize the risk of an unauthorized user accessing bank services. For user-convenient bank, the framing follows Figure 1b) while the neutral scenario decision does not include warning notes.

Table 1 presents a comparison of both framing setups instructions. The readability indices are used to evaluate how easy is to understand the presented text. Readability tests, for instance, Flesch and Kinkaid score [20, 21], are commonly used metric for this purpose. For our instructions, reading ease is 59 % with a study grade level estimated at 7 to 8 grade for framing setup 1 and 6 to 8 grade for the second framing setup. Based on the test, both framing instructions are easy to read for 14 to 16 years old, though framing setup 2 grade level is lower than the obtained for setup 1 for all the scenarios except for forensic. In general, readability index is a good measure of the clarity of the task description though, in the case of crowdsourcing frameworks, brief and simple instructions are considered a good practice [22].

3.4. Datasets

The speaker discrimination sample pairs (trials) are subsets of RedDots [13] and 2018 Voice Conversion Challenge (VCC2018) [10] data. RedDots, one of the widely adopted text-dependent corpora within the ASV community, contains natural speech of short pass-phrases in English, recorded with different smartphones by speakers with diverse native language backgrounds. The trials are content-matched and restricted to same-gender pairs. Additionally, we carefully selected the speaker pairs with matched accents based on the speaker's self-reported native language. The 40 trials correspond to 20 target (same speaker) and 20 non-target (different speaker) comparisons from 12 female and 30 male speakers. The male speakers appear in only one trial, 10 speakers in target and 20 in non-target trials. The samples were normalized to maximum amplitude -1.0 dB with version 2.2.1 of Audacity(R) editing software [23].

The VCC2018 data contains, in addition to natural speech, a powerful voice conversion 'attack' (system N10) to misguide listeners. The trials correspond to read speech from the *device and produced dataset* (DAPS) [24], including four female and four male professional US English speakers. The selected 60 trials correspond to 20 target, 20 non-target and 20 *spoof* trials. The last type of trials correspond to pairs where one sample is from a specific target speaker's natural voice while the other sample is a computer-converted voice of another speaker, with the purpose of sounding the target voice as closely as possible; we point the interested reader to [25] for broad discussion on how voice conversion relates to security considerations of ASV. Due to data limitations, the compared samples, in contrast to RedDots data, do not include content-matched phrases. The non-target and spoof trials are a subset of the similarity listening test of the VCC2018 evaluation [10].

The selected trials from both datasets are gender balanced and the order of the trials was randomized for each scenario. Also the left and right sample order was chosen at random before the tests. The above are aspects to consider in design and implementation of general audio evaluation tests [26].

3.5. Listeners

We utilized Amazon's Mechanical Turk (MTurk[1]) service to perform the listening tests. We required the listeners to have native or advanced English skills and United States as their location. No demographic data was collected from the listeners and none reported hearing problems. Each trial was evaluated by five different listeners. Listeners could answer multiple trials, listen to each trial as many times as needed. Listeners were not informed of the number of target and non-target trials.

In **framing setup 1**, the listeners selected their decisions from a provided 4-point scale following [10]: *Definitely same speaker, same speaker but not sure, different speaker but not sure, definitely different speaker*. We presented the datasets with the corresponding scenario instructions to eight independent panels (two corpora $\times$ four scenarios) for a total partic-

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

Table 1. Readability indices for the instructions in framing setup 1 and 2. Reading ease from a scale [0-100] with higher score for easy to read. Grade level index indicate the years of formal education. Similarly, the readability index (3rd number) estimates the grade level based on the automatic estimation of characters in the text.

	Framing setup 1				Framing setup 2			
Readability Indices	Neutral	Forensic	User-conv. bank	Secure bank	Neutral	Forensic	User-conv. bank	Secure bank
Flesch Kincaid Reading Ease	60.1	61.7	63	67.4	70.1	59.1	69.2	71.2
Flesch Kincaid Grade Level	8.7	8.5	8.7	7.4	6.2	8.2	7	6.7
Automated Readability Index	8	8.4	8.9	7.2	6.86	8	7.5	7.5

ipation of 242 listeners and 2000 decisions distributed in 500 per role play scenario. The instructions (role play scenarios) were the same on RedDots and VCC2018. As the listeners were not informed about spoofing attacks, we expect these trials to strongly influence (spoof) false alarm rates of all the four role play scenarios.

In **framing setup 2**, four independent panels provided decisions with the simplified instructions scenarios for the VCC2018 dataset. The listeners decisions correspond to one of two alternatives: *Definitely same speaker, definitely different speaker*. 180 listeners provided 1200 decision distributed in 300 per scenario. This framing setup intends a more controlled experiment with easy to follow instructions, binary decision, and similar quality in the audio recordings. The VCC2018 dataset contains speech from professional speakers with identical recording conditions in contrast to the RedDots dataset that has a variety of recording devices and diversity in the speakers accents.

4. RESULTS

Graphical summaries of pooled listener responses per scenario and both framing setups are shown in Figures 2 and 3. For framing setup 1, miss rates are relatively low for all the scenarios while false alarms are higher specially for the user-convenient bank scenario. On VCC2018 dataset, false alarms are also high for the forensic scenario. For framing setup 2, the miss rates are relatively low, specially for the user-convenient bank scenario where the false alarm was also higher. Forensic scenario shows the lowest false alarm. The VCC2018 spoof trials show that listeners are easily misguided in both framing setups, as expected.

Summaries concerning miss and false alarm rates are provided in Tables 3 and 4. A sensitive index, d' [27], is included as a measure of discrimination accuracy. d' is defined in terms of the z-score of the hit and miss rates. Comparing the performance between the role play scenarios in framing setup 1, we found that forensic vs. neutral has a higher false alarm rate while the miss rate is similar (RedDots) or relatively higher (VCC2018). The user-convenient bank vs. neutral has higher false alarms in both datasets and the misses are similar in RedDots and higher in VCC2018. Secure bank compared to the other scenarios has the lowest false alarms and misses for the VCC2018, while for RedDots, secure bank's error rates were the second lowest after the neutral scenario. The listeners discrimination accuracy for RedDots data was higher for neutral ($d' = 1.23$), while for VCC2018 was secure bank ($d' = 2.63$). For framing setup 2 (Table 4), forensic has lower false alarm and miss rate than neutral. The user-convenient bank compared to neutral has lower miss and higher false alarm. Secure bank had higher false alarm than neutral, but compare to user-convenient bank it had lower false alarm.

In brief, the results indicate that, while there are clear differences in the observed error rates across different role play scenarios, the direction (increase or decrease) did not correspond to our expectation for framing setup 1. For framing setup 2, there were more changes towards the expected outcome. This may explain the effect of long vs. simplified in-

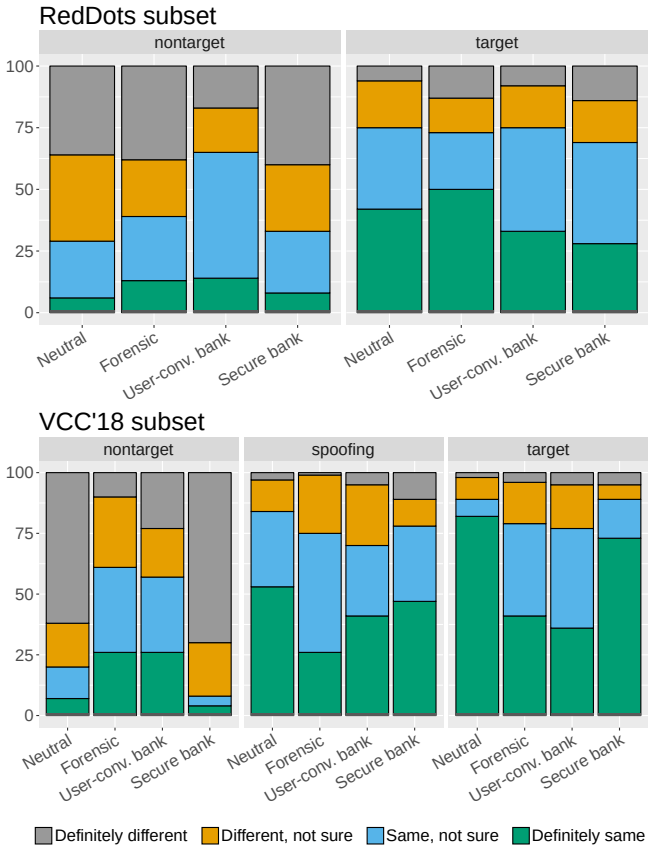

Fig. 2. Listeners' decisions for the selected trials from RedDots and VCC2018 subset presented by evaluated scenario of framing setup 1.

Table 2. Effect of role play scenario on listeners' decisions in terms of the mean difference between the listener panels at 99% confidence interval. Significance codes: '***' 0.001, '**' 0.01, '*' 0.05.

	Target			Non-target			Spoofing	
	Framing setup 1		Framing setup 2	Framing setup 1		Framing setup 2	Framing setup 1	Framing setup 2
	RedDots	VCC18	VCC18	RedDots	VCC18	VCC18	VCC18	VCC18
Forensic – Neutral	-0.01	-0.53 ***	0.18	0.15	1.12 ***	-0.15	-0.34 *	0.06
User-conv. bank – Neutral	-0.11	-0.61 ***	0.36 *	0.63 ***	0.95 ***	0.54 **	-0.28	0.90 ***
Secure bank –Neutral	-0.28	-0.12	0.21	0.02	-0.23	0.24	-0.2	0.66 ***
User-conv. bank – Forensic	-0.1	-0.08	0.18	0.48 *	-0.17	0.69 ***	0.06	0.84 ***
Secure bank – Forensic	-0.27	0.41 **	0.03	-0.13	-1.35 ***	0.39 *	0.14	0.60 ***
Secure bank – User-conv. bank	-0.17	0.49 ***	-0.15	-0.61 ***	-1.18 ***	-0.30	0.08	-0.24

Table 3. Listeners performance in terms of miss (%), false acceptance rate (%) and d' per scenario in framing setup 1.

	RedDots				VCC'18			
	Neutral	Forensic	User-conv. bank	Secure bank	Neutral	Forensic	User-conv. bank	Secure bank
P_{miss}	25.0	27.0	25.0	31.0	11.0	21.0	23.0	11.0
P_{fa} (nontarget)	29.0	39.0	65.0	33.0	20.0	61.0	57.0	8.0
d'	1.23	0.89	0.29	0.94	2.07	0.52	0.56	2.63
P_{fa} (N10 spoof)	n/a	n/a	n/a	n/a	84.0	75.0	70.0	78.0
d' (N10 spoof)	n/a	n/a	n/a	n/a	0.23	0.13	0.21	0.45

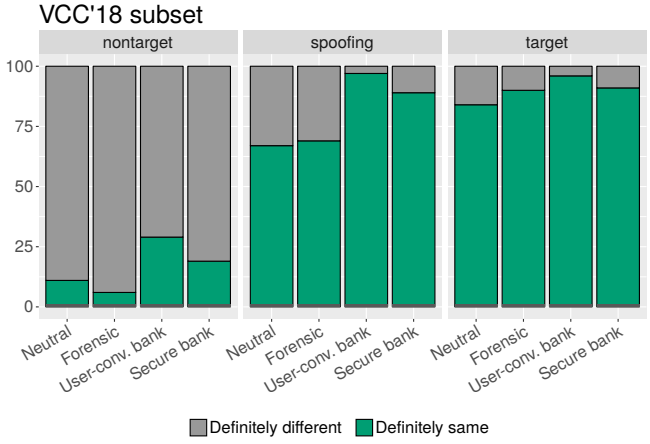

Fig. 3. Framing setup 2 listeners' decisions for VCC 2018 selected trials presented by evaluated scenario.

structions and how the clarity of the task affect the results in crowdsourcing.

A one-way analysis of variance (ANOVA) [28] was further conducted to compare the effect of the role play scenario on the listeners' decisions. Table 2 shows the post hoc pairwise comparison based on the listeners' decisions mean differences using Tukey's honest significance test (HSD) between the scenarios. The level of significance at 99% confidence interval is shown in the corresponding scenario pairs. A significant effect of scenarios on the decisions was found for non-target trials from both framing setups at $p < 0.001$. For target trials, VCC2018 shows a significant difference in decisions due to the scenario effect from framing setup 1.

5. CONCLUSIONS

We addressed the potential impact of speaker discrimination role play scenarios to the decision making of listeners with the help of four role plays. We **confirm** that listener decisions can be influenced: Table 2 showed significant differences in target and nontarget decisions. Concerning our research hypotheses, framing setup 1 did not systematically impact the error we wanted — though, in some cases, we saw evidence of error trade-off behavior. Table 3 for RedDots indicates that while we saw no decrease in P_{miss} from neutral to user-convenient bank, P_{fa} was dramatically increased from 29.0% to 65.0%. This shows some evidence of error trading off. In Framing setup 2, a more controlled setup for the experiment, 4 indicates more favorable results to our hypothesis, with P_{fa} decreased with respect to neutral from 11.0% to 6%, while P_{miss} of user-convenient bank was lower than neutral and other scenarios. Though secure bank has a higher P_{fa} than neutral, the comparison between the secure vs. user-convenient bank shows the trade-off behavior with lowered false alarm and increased miss rate relative to user-convenient bank.

Table 4. Listeners performance (% errors) in terms of miss and false acceptance rate and d' for framing setup 2.

	VCC'18			
	Neutral	Forensic	User-conv. bank	Secure bank
P_{miss}	16.0	10.0	4.0	9.0
P_{fa} (nontarget)	11.0	6.0	29.0	19.0
d'	2.22	2.84	2.30	2.22
P_{fa} (N10 spoof)	67.0	69.0	97.0	89.0
d' (N10 spoof)	0.55	0.79	−0.13	0.11

The different answers between the framing setups hints that shorter and clearer instructions work better in controlling the listeners decisions in each scenarios, specially if the listeners come from crowdsourcing platform.

A possibly relevant consideration would be inclusion of not only *loss*-, but *gain*-targeted role plays. Our preliminary findings suggest potential for substantial further work along these directions.

6. REFERENCES

[1] Gregory Sell, Clara Suied, Mounya Elhilali, and Shihab Shamma, "Perceptual susceptibility to acoustic manipulations in speaker discrimination," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 911–922, 2015.

[2] Robin Panneton Cooper, Jane Abraham, Sheryl Berman, and Margaret Staska, "The development of infants' preference for motherese," *Infant Behavior and Development*, vol. 20, no. 4, pp. 477 – 488, 1997.

[3] Diana Van Lancker and Jody Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[4] Corrina Maguinness, Claudia Roswandowitz, and Katharina von Kriegstein, "Understanding the mechanisms of familiar voice-identity recognition in the human brain," *Neuropsychologia*, vol. 116, pp. 179–193, 2018.

[5] Sarah V Stevenage, "Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings," *Neuropsychologia*, vol. 116, pp. 162–178, 2018.

[6] Astrid Schmidt-Nielsen and Thomas H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Processing*, vol. 10, no. 1, pp. 249 – 266, 2000.

[7] Craig S. Greenberg, Alvin F. Martin, George R. Doddington, and John J. Godfrey, "Including human expertise in speaker recognition systems: report on a pilot evaluation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, p. 5896 5899.

[8] Wade Shen, Joseph Campbell, Derek Straub, and Reva Schwartz, "Assessing the speaker recognition performance of naive listeners using mechanical turk," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5916 – 5919.

[9] Rosa González Hautamäki, Ville Hautamäki, Padmanabhan Rajan, and Tomi Kinnunen, "Merging human and automatic system decisions to improve speaker recognition performance.," in *Proc. Conf. of the Int. Speech Commun. Assoc. (Interspeech)*, 2013, pp. 2519–2523.

[10] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc Speaker Odyssey 2018*, Les Sables D'lonne, France, 2018.

[11] Andrzej Drygajlo, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen, and Tuija Niemi, *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*, Verlag für Polizeiwissenschaft, Frankfurt, Germany, 2015.

[12] Z. Wu, P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, April 2016.

[13] K.A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D.A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, M.J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2996–3000.

[14] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 2010, p. 14.

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

[16] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley, New York, 2 edition, 2001.

[17] E. T. Jaynes, *Probability Theory: The Logic of Science*, p. 589600, Cambridge University Press, 2003.

[18] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty.," *Journal of Risk and Uncertainty*, vol. 5, pp. 297–323, 1992.

[19] Daniel Kahneman and Amos Tversky, "Prospect theory: An analysis of decisions under risk," *Econometrica*, pp. 263–291, 1979.

[20] Rudolf Flesch, "A new readability yardstic," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[21] Georgelle Thomas, R. Derald Hartley, and J. Peter Kincaid, "Test-retest and inter-analyst reliability of the automated readability index, flesch reading ease score, and

the fog count," *Journal of Reading Behavior*, vol. 7, no. 2, pp. 149 – 154, 1975.

[22] Amazon.com, "Requester best practices guide," 2011.

[23] Audacity Team (2019), "Audacity(R): Free audio editor and recorder [Computer application]," Version 2.2.1 retrieved 11 Dec 2017, Audacity(R) software is copyright (c) 1999-2019 Audacity Team.

[24] Gautham J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? - A dataset, insights, and challenges," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1006–1010, 2015.

[25] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130 – 153, 2015.

[26] Soren Bech and Nick Zacharov, *Perceptual audio evaluation. Theory, method and application*, Wiley, England, 2006, ISBN 978-0-470-86923-9.

[27] Neil A. Macmillan and C. Douglas Creelman, *Detection theory: A user's guide*, 2n. ed. Lawrence Erlbaum Associates Publishers, 2005.

[28] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning*, vol. 112, Springer, 2013.