ASR-BASED DEVELOPMENT OF CHALLENGING SPEAKER DISCRIMINATION TESTS

Andrea Fröhlich[1,2,3], Volker Dellwo[2], Peter French[2], Meike Ramon[3]

[1]Zurich Forensic Science Institute, Switzerland
[2]Department of Computational Linguistics (UZH), University of Zurich, Zurich, Switzerland
[3]Applied Face Cognition Lab, University of Lausanne, Switzerland

ABSTRACT

We report our ongoing efforts towards the development of challenging speaker discrimination tests. Our goal is to establish highly sensitive tests that enable characterization of individual differences in voice processing skills. Such tools are required but lacking for the identification of so-called "voice super-recognizers" – individuals with superior voice processing ability – who could aid criminal investigations involving audio material. To find such individuals we manipulated test difficulty using an ASR and delta F0-based stimuli selection method. Independent groups of participants performed 2-alternative forced choice voice discrimination tests of stimulus pairs selected either randomly, or systematically based on inter-item similarity. On average, performance was significantly higher for the prior (80.0%) as compared to the latter (68.8%) test. Thus, we have established a method to manipulate task difficulty in speaker discrimination tests. We will further validate and extend our method to include a wider range of audio material implemented as a within-subject design.

Keywords: voice super-recognizers (VSR), speaker discrimination, forensic phonetics.

1. INTRODUCTION

In criminal proceedings with audio files as evidence, police and prosecution authorities can seek the assistance of forensic phoneticians. Recently, criminal cases with voice evidence have increased where forensic phoneticians are expected to analyse mounting volumes of data in ongoing police investigations. Scenarios such as potentially finding a voice in an extensive collection of case-relevant audio files are impossible to solve by applying traditional auditory-phonetic and acoustic methods within a reasonable time frame. Therefore, new techniques or approaches are required to deal with criminal investigations involving large numbers of audio files. One potential approach to this big data and time-pressure dilemma is the employment of automatic speaker recognition (ASR) systems. They can rapidly calculate similarity scores or (log) likelihood ratios between large numbers of files. However, the performance of these systems is highly affected by the quality and the duration of the audio recordings [1] and humans have also been shown to outperform them in certain scenarios [2]. Therefore, ASR systems are typically used only alongside traditional methods in forensic casework [3]. If, after using ASR systems, forensic phoneticians must still analytically review all the ASR-generated results by applying auditory and acoustic methods, big data and time pressure of ongoing investigations still pose a significant challenge. We are therefore seeking additional solutions.

Since the considerations above are not exclusive to auditory material, ongoing efforts from the domain of vision sciences may offer a solution. For the past decade, researchers have been investigating so-called "Super-Recognizers" (SR), individuals with untrained exceptional face identity processing abilities [4],[5]. As interest from international law enforcement agencies steadily increases [5], [6], so do reports of formal collaborations between researchers and police agencies who seek to actively deploy individuals due to an empirically documented unique ability [5].

Notwithstanding that SR are a very recent phenomenon in the visual domain, we have adopted an interdisciplinary approach to determine whether an analogous superior ability also exists for the processing of voice identity. Our current definition of "voice super-recognizers" (VSR) describes untrained individuals with *superior ability to efficiently process speaker identity consistently across input variations*. This could manifest as highly proficient speaker discrimination and/or recognition across objectively challenging conditions associated with increased task difficulty. Such conditions could include, e.g., low-quality audio material or high similarity of speakers' voices. Following this definition, we currently envision the employment of VSR in post-processing of ASR-generated results in police investigations or intelligence gathering with big data cases (Figure 1).

Figure 1: Potential employment scenario of VSR.

Given the nature of to-be-processed material, we recommend identifying VSR among police employees.

One approach to identify VSR proposed by Jenkins *et* al. [7] is to test if the abilities of individuals who identified as visual SR, also extend to voice processing.

Their participants completed the Bangor Voice Matching Test [8] (discrimination test), the Glasgow Voice Memory Test (recognition test) [9] and a bespoke Famous Voice Recognition Test [7]. The authors reported positive correlations between voice- and face processing and that some but not all individuals with superior face ability showed superior performance for voice processing.

We sought to expand on this work and further explore the concept of VSR by developing tests incorporating a larger quantity of speaker identities while systematically varying task difficulty using automatic approaches.

The current study proposes a method to control task difficulty by selecting voice stimuli based on within- and between-speaker similarity. We created a speaker discrimination test based on ASR-generated similarity scores, analogous to how we envision the employment of VSR to post-process ASR results. This provides the basis for creating a challenging test of speaker discrimination, which is the focus of the present work.

1.1 Previous work

To the best of our knowledge, the subject of *superior* voice processing ability, an umbrella term for various voice and speaker-related tasks, has been addressed in only a small number of studies so far [8]–[10]. Mühl *et al.* developed the aforementioned Bangor Voice Matching Test to assess voice discrimination abilities in a short and standardised screening test. While their research on speaker discrimination focused on assessing listeners' abilities, other groups focused on factors influencing task difficulty [11]. Alongside 'natural' variations, such as speaking style that render speaker discrimination tasks more challenging, acoustic properties of voices and their correlation to listeners' judgments in speaker discrimination tests have been investigated [12]. While speaker discrimination tests force listeners to make binary decisions, similarity ratings give more granular insights into the perceived similarity scale of voices [13]. As to the correlation between automatic and human similarity ratings, so far only phonetic features were investigated [14]. As far as we know, there is currently no research indicating that MFCC-based ASR systems correlate with human speaker discrimination performance. Still, we suspect that stimulus pairs rated similarly by ASR systems will be harder to discriminate by humans.

In the domain of human assisted speaker recognition (HASR), it was found that fusing machine-generated and human results showed improvements [2], [15], [16]. These findings provided inspiration for how VSR could be employed in investigative case work for post-processing ASR system-generated results. Contrary to some of the analysed schemes, we envision the employment of lay people only in the investigatory stage of cases.

2. METHOD

Different factors influence the difficulty of a speaker discrimination task. For this test, we only investigated and tested "voice-inherent factors", such as two voices sharing similar frequency properties [17], by using ASR-derived similarity scores and delta F0 values to find challenging stimulus pairs with low within-speaker, and high between-speaker similarity. We have refrained from making the experiment more difficult by addressing "technical factors" (such as recording devices) other than re-sampling the stimuli down to 8 kHz. For future experiments, we also plan to implement "acoustic environment factors", such as background noise, reverberation etc. We only tested short stimuli of around 1.2 seconds as Bricker and Pruzansky have shown that participants' discrimination accuracy increased with longer speech sample duration [18].

2.1 Corpus and participants

We used stimuli from the TEVOID corpus (Temporal Voice Idiosyncrasy, [19]), initially created to investigate speakers' temporal features in short sentences. We further processed the studio quality recordings that consist of read sentences by native Zurich German speakers (male and female) as described below. The study was run with participants from different classes of police cadets in Zurich, Switzerland.

2.2 Stimulus selection

As in a possible scenario where VSR would assist in a real police case, we combined an ASR system to pre-process the data for the speaker experiment. The goal of using an ASR system was to have a fast method to select similar stimuli from large corpora.

Yet, most ASR systems show high equal error rates with short speech sequences since there is insufficient information to build an adequate speaker model [2], [20]. To mitigate this challenge, we worked with a state-of-the-art speaker comparison model for short stimuli, which won the "Short-duration Speaker Verification Challenge 2020" [21],[22]. This ECAPA-TDNN algorithm (release 05-03-21) is open-source from Speechbrain [23] (version 0.5.10) and was included in our stimuli processing pipeline implemented in Python (version 3.8.12).

To select the stimulus pairs for the discrimination experiment, we first loaded all stimuli from the corpus using the `AudioNormalizer` and created embeddings using the `EncoderClassifier` classes of Speechbrain [23]. We then calculated pairwise cosine distances between the embedding vectors to obtain a full similarity matrix for all stimuli. We use the pairs with the lowest within-speaker and the highest between-speaker similarity scores. No likelihood ratios were calculated, the system was solely used on a score basis.

It is noteworthy that in the MFCC-feature extraction process, F0 information is neglected [21],[22]. However, it has been found that mean pitch influences listeners' judgements of voice dissimilarity [17]. We therefore measured and averaged the fundamental frequency (F0) of all audio samples for each speaker and calculated "delta F0" values for all speaker pairs. After investigating the F0 distribution of male and female voice identities in TEVOID, we only considered stimulus pairs for speakers with a delta F0 of 10 Hz or less for female speakers and 15 Hz for male speakers.

We then sorted all stimulus pairs by ascending cosine similarity and started picking pairs from the top of this list. A maximum number of four stimuli per speaker were included for different-speaker and same-speaker scenarios each. Two speakers only appeared in a pair with each other up to two times. The two stimuli of a pair never originated from the same sentence, and each selected audio stimulus only appeared once in the whole experiment. We performed all the above processing steps separately for male and female speakers and used ten speakers per gender. Overall, the experiment included 20 same-speaker (SS) and 20 different-speaker (DS) trials per gender.

2.3 Stimulus processing

To harmonise the audio file durations, which varied inter-individually with speakers' speech tempo, we normalised the selected stimuli by cutting the files to a duration of around 1.2 seconds, chosen from the centre of the audio file. We normalised the amplitude to 65 dB (RMS), but not the number of syllables contained in one snippet. For post-processing the selected files, we applied an amplitude smoothing function with the software Praat [24] to make the starting and ending of the stimuli less abrupt. In addition, we removed non-speech sections before and after the utterances.

Furthermore, we down-sampled the resulting stimuli to 8 kHz to make them more similar to a frequently observed setting in forensic casework (i.e. telephone transmitted speech). Finally, we concatenated each stimulus pair in random order with a pause of one second between the stimuli and saved them as MP3 files.

To validate our ASR and delta F0-based stimuli selection method, we implemented a second discrimination experiment where participants were presented with randomly selected stimulus pairs of the same audio quality, length, and out of the same TEVOID-corpus. We followed the same rules concerning the number of files per speaker and the spoken sentences. This random discrimination test was assessed with two unique groups of police cadets.

2.4 Experimental design

We used the Gorilla Experiment Builder [25] to create and host the speaker discrimination experiment. Pairs of audio stimuli (number of pairs = 80) were presented sequentially to the participants in random order, who had to indicate whether both were uttered by the same speaker or by two different speakers. We ran the two experiments on different classes of police cadets on different days. Participants completed the experiment onsite. They all used the same headphones.

3. RESULTS

We initially assessed the discrimination performance of the ECAPA-TDNN model on the original stimuli of the complete TEVOID corpus. The model showed a ROCCH EER [26] of only 0.92% for males and 2.13% for female stimuli. We compared this EER to the one of a x-vector-based system validated for forensic case work that showed a ROCCH EER of 16.1% for the males and a ROCCH EER of 16.5% for the female stimuli.

The results of the speaker discrimination tests are as follows: In the randomly combined stimulus pairs, participants, on average, correctly classified 80.0% of the pairs (38 participants with 80 stimulus pairs each), whereas in the challenging discrimination task, participant groups only classified 68.8% of the pairs correctly (52 participants with 80 stimulus pairs each). This observed difference is significant ($p<0.01$, two-sided t-test). The distributions of correct

answers per experiment are visualised in Figure 2. The challenging task was run with three groups of police cadets, of which the first scored 68.1%, the second 69.1% and the third 70.3% of the pairs correctly.

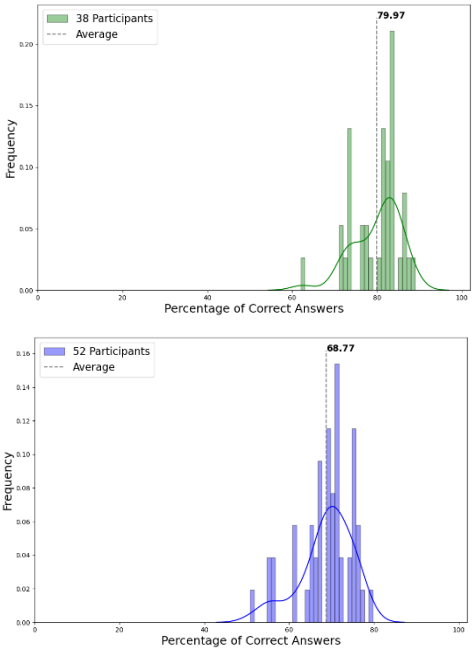

Figure 2: Test result of random (top) and score-based (bottom) stimuli selection method.

4. DISCUSSION

The presented study aimed to implement and test a new stimulus selection method to control voice-inherent speaker similarity in a discrimination task. We did so by employing an ASR model combined with delta F0 values for objective stimulus selection. The used ECAPA-TDNN model was not yet tested on forensic data sets or in a forensic casework scenario as suggested by Morrison and Enzinger [27]. Yet, it performed best in short speech stimuli discrimination and showed very low EER on our stimuli. Also, we did not employ the model to perform any forensic speaker discrimination but only to select difficult stimulus pairs to be used in our test. In future experiments, we will calibrate the ASR-system and try to make the stimuli as similar as possible to a forensic scenario.

So far, we ran our experiment on three groups of police cadets that scored similarly regardless of group size or composition. To further validate our method, we implemented and tested a second experiment with randomly chosen stimulus pairs which was tested on two additional independent participant groups. When comparing the results between random stimulus selections and stimulus selections based on our new method, the test designed using our method turned out to be significantly more challenging. This finding supports our hypothesis that discrimination task difficulty can be controlled by selecting stimulus pairs based on ASR system scores and delta F0 values.

The current test did not yet show any distinct superior group of participants in the distribution of scores. However, one could argue that the participants at the high end of a distribution of voice discrimination and/or recognition abilities could possibly be VSR. VSR capabilities could manifest as highly proficient speaker discrimination and/or recognition across objectively challenging conditions associated with increased task difficulty. Whether the best performers of our test are in fact VSR, will have to be evaluated after they have completed additional challenging and forensically relevant tests. However, we consider a challenging discrimination test like ours a necessary precondition for identifying VSR if they do, in fact, exist.

VSR could pave the way for human-centred approaches in biometrics and forensics with implications for law enforcement. Despite currently non-existent research regarding VSR, their discovery and employment in actual cases receive interest in the community. Currently, we can envision VSR deployment for postprocessing of automatically generated results in investigatory stages of a case involving large amounts of audio files. Identification of VSR will ultimately provide the needed basis for further research, including the benefits of their employment for criminal investigation, and the neural basis of their abilities. Finally, fusing the abilities of VSR and state-of-the art ASR systems could further improve voice processing performance.

We plan to repeat our experiments with more participants and as a within-subject design. Alongside voice-inherent factors, technical and acoustic-environmental factors will further be included in future tests, where we plan to extend our method beyond speaker discrimination to further test scenarios, such as one-to-many speaker recognition, sorting, or clustering tasks. Finally, our upcoming tests will include forensically relevant scenarios and stimuli to potentially find VSR best equipped for the employment in investigatory case work.

7. REFERENCES

[1] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, and A. Alexander, 'Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)', Speech Commun., vol. 112, pp. 30–36, Sep. 2019, doi: 10.1016/j.specom.2019.06.005.

[2] S. J. Park, A. Afshan, J. Kreiman, G. Yeung, and A. Alwan, 'Target and Non-target Speaker Discrimination by Humans and Machines', in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 6326–6330. doi: 10.1109/ICASSP.2019.8683362.

[3] E. Gold and P. French, 'International Practices in Forensic Speaker Comparisons: Second Survey', Int. J. Speech Lang. Law, vol. 26, no. 1, pp. 1–20, Jun. 2019, doi: 10.1558/ijsll.38028.

[4] R. Russell, B. Duchaine, and K. Nakayama, 'Super-recognizers: People with extraordinary face recognition ability', Psychon. Bull. Rev., vol. 16, no. 2, pp. 252–257, Apr. 2009, doi: 10.3758/PBR.16.2.252.

[5] M. Ramon, 'Super-Recognizers – a novel diagnostic framework, 70 cases, and guidelines for future work', Neuropsychologia, vol. 158, p. 107809, Jul. 2021, doi: 10.1016/j.neuropsychologia.2021.107809.

[6] M. Ramon and S. Rjosk, beSure? – Berlin Test for Super-Recognizer Identification: Part I: Development. Verlag für Polizeiwissenschaft, 2022.

[7] R. Jenkins et al., 'Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks', 2020, doi: 10.31234/osf.io/7xdp3.

[8] C. Mühl, O. Sheil, L. Jarutytė, and P. E. G. Bestelmeyer, 'The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability', Behav. Res. Methods, vol. 50, no. 6, pp. 2184–2192, Dec. 2018, doi: 10.3758/s13428-0.

[9] V. Aglieri, R. Watson, C. Pernet, M. Latinus, L. Garrido, and P. Belin, 'The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices', Behav. Res. Methods, vol. 49, no. 1, pp. 97–110, Feb. 2017, doi: 10.3758/s13428-015-0689-6.

[10] D. Humble, S. R. Schweinberger, A. Mayer, T. L. Jesgarzewsky, C. Dobel, and R. Zäske, 'The Jena Voice Learning and Memory Test (JVLMT): A standardized tool for assessing the ability to learn and recognize voices', Behav. Res. Methods, Jun. 2022, doi: 10.3758/s13428-022-01818-3.

[11] A. Afshan, J. Kreiman, and A. Alwan, 'Speaker discrimination performance for "easy" versus "hard" voices in style-matched and-mismatched speech', J. Acoust. Soc. Am., vol. 151, no. 2, pp. 1393–1403, 2022.

[12] J. Kreiman, S. J. Park, P. A. Keating, and A. Alwan, 'The relationship between acoustic and perceived intraspeaker variability in voice quality', in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[13] K. McDougall, 'Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades', Int. J. Speech Lang. Law, vol. 20, no. 2, pp. 163–172, Dec. 2013, doi: 10.1558/ijsll.v20i2.163.

[14] L. Gerlach, K. McDougall, F. Kelly, A. Alexander, and F. Nolan, 'Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features', Speech Commun., vol. 124, pp. 85–95, Nov. 2020, doi: 10.1016/j.specom.2020.08.003.

[15] R. G. Hautamäki, V. Hautamäki, P. Rajan, and T. Kinnunen, 'Merging human and automatic system decisions to improve speaker recognition performance', in Interspeech 2013, ISCA, Aug. 2013, pp. 2519–2523. doi: 10.21437/Interspeech.2013-422.

[16] R. Schwartz et al., 'USSS-MITLL 2010 human assisted speaker recognition', in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2011, pp. 5904–5907. doi: 10.1109/ICASSP.2011.5947705.

[17] T. K. Perrachione, K. T. Furbeck, and E. J. Thurston, 'Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices', J. Acoust. Soc. Am., vol. 146, no. 5, pp. 3384–3399, Nov. 2019, doi: 10.1121/1.5126697.

[18] P. D. Bricker and S. Pruzansky, 'Effects of stimulus content and duration on talker identification', J. Acoust. Soc. Am., vol. 40, no. 6, pp. 1441–1449, 1966.

[19] V. Dellwo, A. Leemann, and M.-J. Kolly, 'Speaker idiosyncratic rhythmic features in the speech signal', Interspeech Conference Proceedings, 2012.

[20] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, 'Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles', J. Acoust. Soc. Am., vol. 144, no. 1, pp. 375–386, 2018.

[21] J. Thienpondt, B. Desplanques, and K. Demuynck, 'The IDLab Short-duration Speaker Verification Challenge 2020 System Description', Tech. Rep., 2020.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, 'Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification', ArXiv Prepr. ArXiv200507143, 2020.

[23] M. Ravanelli et al., 'SpeechBrain: A general-purpose speech toolkit', ArXiv Prepr. ArXiv210604624, 2021.

[24] P. Boersma, 'Praat, a system for doing phonetics by computer', Glot Int, vol. 5, no. 9, pp. 341–345, 2001.

[25] A. Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, 'Gorilla in our midst: An online behavioral experiment builder.', 2020.

[26] N. Brümmer and E. de Villiers, 'The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF'. arXiv, Apr. 10, 2013. Accessed: Dec. 31, 2022. [Online]. Available: http://arxiv.org/abs/1304.2865

[27] G. S. Morrison and E. Enzinger, 'Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction', Speech Commun., vol. 85, pp. 119–126, 2016, doi: 10.1016/j.specom.2016.07.006.