Human spoofing detection performance on degraded speech

Camryn Terblanche[1], *Philip Harrison*[2,3], *Amelia Gully*[2]

[1]Department of Speech and Language Pathology, University of Cape Town, South Africa
[2]Department of Language and Linguistic Science, University of York, UK
[3]J. P. French Associates, York, UK

c.terblanche04@gmail.com, philip.harrison@york.ac.uk, amelia.gully@york.ac.uk

Abstract

Over the past few years attention has been focused on the automatic detection of spoofing in the context of automatic speaker verification (ASV) systems. However, little is known about how well humans perform at detecting spoofed speech, particularly under degraded conditions. Using the latest synthesis technologies from ASVspoof 2019, this paper explores human judgements of speech authenticity by considering three common channel degradations – a GSM network, a VoIP network, and background noise – in conjunction with varying synthesis quality. The results reveal that channel degradation reduces the size of the perceptual difference between genuine and spoofed speech, and overall participants correctly identified human and spoofed speech only 56% of the time. In background noise and GSM transmission, lower-quality synthetic speech was judged as more human, and in VoIP transmission all speech, including genuine recordings, was judged as less human. Under all conditions, state-of-the-art synthetic speech was judged as human, or more human than, genuine recorded speech. The paper also considers the listener factors which may contribute to an individual's spoofing detection performance, and finds that a listener's familiarity with the accents involved, their age, and the audio equipment used for playback, have an effect on their spoofing detection performance.

Index Terms: spoofing detection, degraded speech, human performance

1. Introduction

The use of automatic speaker verification (ASV) systems continues to grow across a range of applications including banking, smart devices, and forensics [1]. An area of concern with the technology is the potential for systems to be subverted via spoofing attacks [2]. A spoofing attack usually involves the presentation of an electronically generated, modified or replayed speech signal to an ASV system in an attempt to gain access to a physical location, device or service.

Over recent years, increased attention has been paid to the susceptibility of ASV systems to spoofing attacks and their automatic detection through challenges such as the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenges [3]. These challenges drive research and innovation in the area by providing participants with training and test recordings made using a variety of methods available to would-be attackers.

One related issue which has received relatively little attention concerns how good human listeners are at detecting spoofing attacks. This is of interest from two perspectives. Firstly, as a comparison with automatic systems and secondly, as humans may also encounter real-world situations where they are presented with spoofed speech. This could be in the context of an interaction involving spoofed speech, such as a telephone conversation [4], as the consumer of deepfaked content, or when conducting a forensic speaker comparison examination [5]. As the perceptual quality of speech synthesis and voice conversion technologies improves over time, and they become more widely available, the risk of their use going undetected increases.

This concern is supported by the findings of a recent test of human spoofing detection performance [6], which led to the conclusion that "state-of-the-art TTS has the capability of producing synthetic speech that is perceptually indistinguishable from bona fide speech". When considering real-world conditions, the results of an earlier study [7] showed that human detection performance was worse when the audio signal bandwidth was reduced from 8 kHz to 4 kHz. However, the listeners were not deceived to the same extent as the subjects in [6], most likely due to the less convincing synthetic speech generated by the technologies of the time.

The current study aims to provide further insights into the spoofing detection performance of human listeners by using a subset of the synthesis technologies tested in [6] and degrading the speech to simulate three channel conditions under which spoofed speech may be encountered in the real-world. The study also investigates to what extent listener factors affect performance. Given the findings of [7], we predict that the channel degradations will lead to a decrease in performance in identifying both genuine and synthesised speech. The remainder of this paper is laid out as follows: Section 2 describes the experimental design, Section 3 presents results which are then discussed in Section 4, and conclusions are presented in Section 5.

2. Methods

The aim of this study is to compare synthesis quality, common types of channel degradation, and listener factors on human spoofing detection performance. A variety of audio samples were selected and processed as described below. Both male and female voices were used for each condition to provide a repeat trial. A total of 32 samples were presented to each listener (two samples of every possible combination of synthesis quality and channel degradation).

2.1. Synthesis quality

Speech from the ASVspoof challenge 2019 [3] was used in this study. A number of text-to-speech (TTS) systems from the ASVspoof 2019 challenge were compared in [6], and it was

http://dx.doi.org/10.21437/Interspeech.2021-1225

found that their perceptual quality varied. This study makes uses of three synthesis systems identified in [6], covering the range of possible qualities of spoofing attack:

- A08 (lowest perceptual synthesis quality). A NN-based TTS system. A08 uses a neural-source-filter waveform model. Attackers may use this system if they want to generate fake speech at a high speed [6].
- A07 (medium perceptual synthesis quality). A NN-based TTS system. It may be used by attackers if they intend to leverage the GAN-based post-filter, with the hope that the GAP filter may mask differences between the generated speech waveform and natural speech waveform [6].
- A10 (highest perceptual synthesis quality). An end-to-end NN-based TTS system that applies transfer learning from speaker verification to a neural TTS system called Tacotron 2. Attackers may use A10 as it is reported that synthetic speech produced by this system has high naturalness and good similarity to target speakers perceptually [6], [8].
- Genuine speech. The clean full-bandwidth (16 kHz sample rate) audio file.

2.2. Channel degradation

Samples were also processed to simulate each of the following common channel degradations:

- Internet video conference call. A Zoom call was made between two computers, each using the Zoom desktop client (version 5.0.4) [9]. A sound file containing multiple repetitions of the source material was replayed on one computer via the screen sharing function, with 'share computer sound' enabled. The network connection of the receiving computer was artificially degraded using Clumsy [10] with an 85% chance of incoming frames being dropped. This resulted in distortions to the speech caused by the dropped frames. The material was recorded on the receiving computer and instances were selected that contained obvious degradations.
- GSM mobile telephone. The GSM AMR Codec Platform [11] was used to simulate a mobile telephone call made on a GSM network. The lowest bitrate of 4.75 kbps (constant) was used to simulate a poor-quality connection.
- Background noise. A recording of background noise in a crowded restaurant [12] was mixed with the speech samples at a suitable level to mimic a real-life conversation in such an environment.
- Clean audio with no degradations.

2.3. Test design

An online listening test was set up in Qualtrics [13]. Online delivery offered access to a much greater participant pool, but also resulted in a less controlled testing environment than an in-person listening test. The test consisted of four sections: a biographical question stage, a familiarization stage where participants could practice answering the questions for two samples, the main testing stage, and a debrief stage.

During the biographical question stage, participants were asked questions about their age, sex, level of English language proficiency, linguistic training, country of socialization, and familiarity with VoIP and telephone calls, in order to determine whether any of these factors were relevant to the spoofing detection task.

During the main testing stage, a role-play scenario was presented in which the participant was asked to imagine that they were responsible for detecting spoofing attacks on a banking system, and they must detect which speech samples were made by humans. Their decision was recorded using a Likert scale from 1 (definitely synthesized) to 7 (definitely human). As a result, participants were aware that spoofed speech was present in the samples. Sample presentation order was randomized, but participants were told whether the sample was being received over the internet, telephone, or in a noisy restaurant. Participants could listen to each sample as many times as they wanted, but the back button was disabled. The number of times each clip was played, and participant response times, were recorded. All audio in the test was presented as 44.1 kHz mono wav files for maximum compatibility with presentation over internet browsers.

During the debrief phase, participants were asked about the audio equipment they used and environment they were in when completing the task (participants were encouraged from the start to use headphones in a quiet environment, but this was not possible for all participants), how confident they were in their answers overall, and whether they experienced any technical difficulties.

In total, 179 participants (mean age 34, age range 18-67) from 25 different countries completed the survey, with 75% having English as their first language. Average completion time was 44 minutes, with an average of 1.67 listens per sample. Generalized linear mixed-effect models fit by maximum likelihood (hereafter `glmer`) were conducted using R [14] with the lme4 package [15]. Responses were converted to binary values (human / not human) and 161 instances where the middle option on the Likert scale was selected were removed from the analysis. 14 participants were removed from the analysis (6 due to self-reported hearing difficulties, 8 due to not listening to all samples before responding), and data from the remaining 165 participants were analyzed.

3. Results

3.1. Synthesis quality and channel degradation

Figure 1 illustrates the effect of channel quality (CQ) and synthesis quality (SQ) on listener perception. The `glmer` model reported that the effect of the interaction between CQ and SQ was significant ($\chi^2(9) = 50.43$, $p<.05$). It is clear from Fig. 1 that listener perceptions of spoofed speech vary substantially depending on both SQ and CQ, and that the highest perceptual quality synthesized speech (A10) was rated as comparable to, or more human than, genuine recorded speech in all conditions, whether clean or degraded. In general, the effect of adding channel degradation was to reduce the size of the perceptual difference between genuine and the poorer quality spoofed speech (A08 and A07).

For clean audio (Fig. 1(a)), the performance was comparable for the low and medium SQ speech A08 and A07, as they were only mistaken for human around 20% of the time. On the other hand, high SQ speech A10 was directly comparable with genuine recorded speech, both being perceived as human around 80% of the time. The variance for all SQs is roughly the same at around ±5%, indicating a good degree of listener agreement.

In background noise, there was a rise in the perceptual acceptability of A08 and A07 to around 45%. The genuine

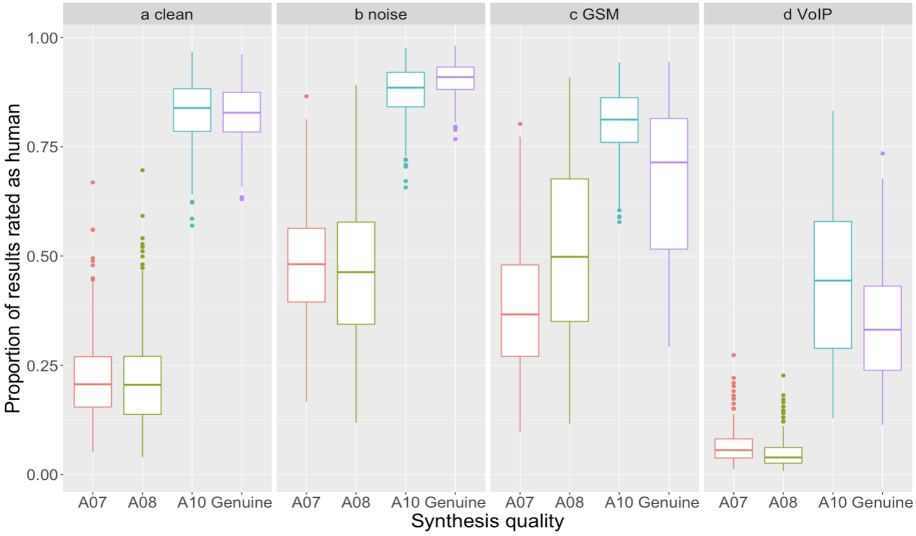

Figure 1: *Proportion of sound files per listener rated as human, for a) the clean speech signal, b) speech signal in background noise, c) speech signal with simulated GSM transmission, d) speech signal with simulated VoIP transmission.*

speech was identified as human around 90% of the time with very good listener agreement, and high SQ speech A10 was again judged similarly to the genuine speech with a median of 88% human.

In GSM transmission, the perceptual acceptability of A08 and A07 is similar to that in background noise, although interestingly the medium perceptual SQ (A07) performs worse than the lowest (A08), albeit with lower listener agreement. High SQ speech A10 was rated as most human in general, with good listener agreement. Perhaps surprisingly, given the prevalence of mobile phone transmission, genuine speech was rated less human in general than A10, but with a large amount of listener variation.

Finally, in simulated VoIP transmission, perceptual ratings for all SQs were considerably lower than in other channel conditions, perhaps due to the more artificial-sounding nature of VoIP degradation and frame/packet loss. The performance of A08 and A07 are particularly poor for this CQ, with very good listener agreement, but the median rating for all SQs is under 50% human in this condition. As with GSM transmission, the high SQ speech A10 is rated more human overall than the genuine speech.

Given the wide variance in results for some conditions, it appears that some listeners are better at detecting spoofed speech in degraded conditions than others. In order to explore this further, listener factors were also analyzed.

3.2. Listener factors

Figure 2 illustrates the effect of selected listener factors on spoofing detection performance. The glmer model shows a significant effect of country of socialization on the accuracy of authenticity judgements ($\chi^2(2) = 11.79$, p= .003), shown in Fig. 2(a), with participants from the UK 1.35 times more likely to be correct.

Figure 2(b) displays the effect of the significant interaction between SQ and audio equipment ($\chi^2(3) = 31.03$, $p<.05$). The 45% of participants who listened via headphones, rather than speakers, achieved better results when identifying the low and medium SQ and were more consistent in their responses across the range. However, listening to the good SQ with speakers led to increased accuracy.

Figure 2(c) shows the effect of the significant interaction between CQ and age ($\chi^2(6) = 13.59$, p=.035). There was greater variation between participants when the CQ was manipulated compared to clean audio. GSM transmission and background noise produced the poorest results. Across all age groups, most correct responses were given in the VoIP condition, and young adults (ages 18-34) performed best. Conversely, for GSM transmission, young adults performed worst, and older adults (ages 51 and over) performed best.

4. Discussion

The main hypothesis of this study was that human spoofing detection accuracy would decrease as synthesis quality increased and as channel quality decreased. The results support this hypothesis, with the best quality synthesis consistently being judged as more human than genuine speech, and with the perceptual distance between synthetic and genuine speech decreasing in the presence of channel degradation and noise.

In the presence of background noise, it was expected that the speech signal would be somewhat masked, resulting in reduced spoofing detection accuracy. The results bear this out, and interestingly, all speech qualities were rated as more human in background noise than in the clean channel, suggesting an effect of the background noise on the perception of even genuine speech. Restaurant noise was selected to present a realistic scenario, but since it comprises largely of speech noise, it may act as an informational as well as an energetic masker [16]. It would be interesting to perform a similar study comparing the effect of informational and purely energetic maskers (such as white noise) on spoofing detection performance.

When VoIP channel degradation was simulated, all speech signals were judged to be more synthetic. This may indicate that the type of interference expected by listeners primed to detect spoofing attacks corresponds to the type of interference produced in a VoIP channel, leading to lower scores overall.

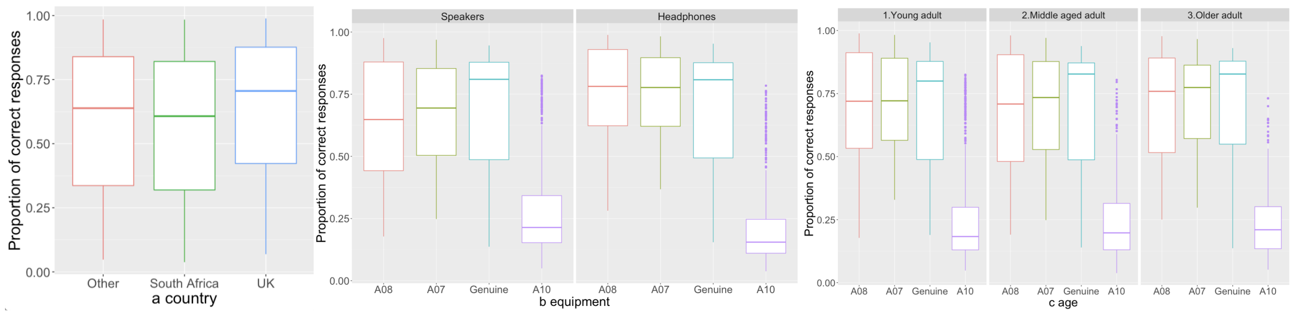

Figure 2*: Effect of listener and equipment factors on the proportion of correct responses per listener by: a) country of socialization, b) playback equipment and c) age group.*

The lower SQ systems (A07 and A08) obtained the lowest overall scores in the VoIP condition, whereas the highest SQ system was rated as more human than genuine speech, suggesting an interaction between SQ and CQ for this channel. In a simulated GSM channel, the effect on synthetic speech was similar to adding background noise, but the genuine speech was rated as less natural, suggesting that spoofing detection in a GSM channel is unreliable.

In all degraded channel conditions, the perceptual distance between genuine speech and the lowest-rated synthetic speech was decreased relative to clean audio. It appears that whatever features listeners use to distinguish between genuine and spoofed speech lose some of their discriminatory potential in the presence of channel degradation.

For all channels, the highest SQ sample was rated as human as, and in some cases more human than, genuine speech. It is therefore clear that speech synthesis technology has improved to such an extent that critical listening alone is no longer sufficient to determine authenticity, particularly when the CQ emulates realistic conditions. The lowest and medium SQ samples were generally rated approximately equally, indicating that in degraded conditions, the difference between the SQs is less perceivable than it was found to be in [6].

Of the listener factors investigated, it was found that a listener's country of socialization – and thereby their familiarity with the English and Scottish accents used in the test – influenced their spoofing detection performance. It follows that, based on research into the other-accent effect [17, 18], accent familiarity may be an indicator of one's ability to judge the authenticity of speech.

Older adults showed slightly better performance at authenticating mobile calls than younger listeners. This may be because they defer to mobile phone calls rather than other communication tools. However, younger adults are likely to use more modern tools more intensively, with greater exposure to VoIP-like encoded speech [19], which may explain why they are better at authenticating VoIP speech than older participants. The effect between performance and listeners' time spent making internet calls was considered, but the results were not significant. Additionally, in cases where background noise was added, it may have been easier for young adults to selectively pay attention and discriminate the target voice [20]. Further research to explore the effect of age on speech authenticity judgements is recommended.

Perhaps surprisingly, it appears participants who used speakers rather than headphones were better at detecting that good SQ synthetic speech was synthetic. Further research to determine why this occurred is required, but it likely relates to the acoustic characteristics of the playback environment and its interaction with the characteristics of the synthetic speech.

Finally, some general trends were visible in the results which did not reach statistical significance. Individuals who listened to the audio the fewest times produced the most correct scores, suggesting that initial impressions are often correct. It also indicates that highly diagnostic information can be found at the segmental or suprasegmental level, despite the short length of the samples. The final question of the online test asked listeners to rate their overall level of confidence in their responses, and participants who were absolutely sure that they were not fooled seemed to be the least accurate. This finding is in line with previous speaker identification studies and further demonstrates that confidence is not a good predictor of accuracy [17, 21]. It was expected that linguists would outperform naïve listeners, and the trend found suggests that linguists do display slightly higher levels of accuracy. Some individuals were highly accurate, even under degraded conditions, which shows the variability of performance across listeners. However, not every participant is equally well able to discriminate between synthetic and genuine voices, just like not every listener in a close social network is equally able to identify voices [22].

5. Conclusions

This study corroborates [6] as it is clear that state-of-the-art TTS systems have the capability to produce synthetic speech that is perceptually indistinguishable from genuine speech by human listeners. The study suggests that the CQ affects individuals' judgements of authenticity to such an extent that even bad SQ can sometimes prove successful at fooling listeners. Despite being aware that spoofed speech was present, many participants were still fooled, which is a further demonstration of the realism of the latest speech synthesis technologies.

Future work would include conducting similar tests in a more controlled environment, to remove the effects introduced via the lack of control over equipment. More research is required to determine if a full auditory and acoustic analysis of synthesised speech, conducted by trained linguists, would be sufficient to accurately determine authenticity. It may be beneficial to explore the kind of implications that this technology might have in forensic speaker comparison case work, particularly when good quality synthetic material is used. Additionally, the current study highlights human performance on the task of spoofing detection in degraded conditions, but it would be interesting to investigate the correlation between human participants' scores and the scores from automatic-spoofing detection systems when the CQ is manipulated to a similar degree.

References

[1] D. Watt, P. Harrison, and L. Cabot-King, ‘Who owns your voice? Linguistic and legal perspectives on the relationship between vocal distinctiveness and the rights of the individual speaker’, *The International Journal of speech, Language and the Law*, vol. 26, no. 2, pp. 137–180, 2019, doi: https://doi.org/10.1558/ijsll.40571.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, ‘Spoofing and countermeasures for speaker verification: A survey’, *Speech Communication*, vol. 66, pp. 130–153, 2015.

[3] ASVspoof consortium, ‘ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan’. 2019, [Online]. Available: http://www.asvspoof.org/.

[4] C. Stupp, ‘Fraudsters used AI to mimic CEO’s voice in unusual cybercrime case’, *Wall Street Journal*, 2019.

[5] P. Foulkes and P. French, ‘Forensic speaker comparison: a linguistic-acoustic perspective’, in *The Oxford Handbook of Language and the Law*, L. Solan and P. Tiersma, Eds. Oxford: Oxford University Press, 2012, p. The Oxford Handbook of Language and the Law, Chapter 41.

[6] X. Wang *et al.*, ‘ASVspoof 2019: A large-scale public database of synthetic, converted and replayed speech’, *arXiv*, 2020, [Online]. Available: https://arxiv.org/pdf/1911.01601.pdf.

[7] M. Wester, Z. Wu, and J. Yamagishi, ‘Human vs machine spoofing detection on wideband and narrowband data’, in *Paper presented at Interspeech, Dresden, Germany*, Dresden, Germany, 2015, pp. 2047–2051.

[8] J. Shen *et al.*, ‘Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions’, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[9] E. S. Yuan, *Zoom Video Communications*. Zoom Video Communications, 2020.

[10] C. Tao, *Clumsy 0.2 [computer software]*. 2018.

[11] Forensics Speech Research Group, *GSM AMR Codec Platform [computer software]*. University of Auckland, New Zealand: University of Auckland, New Zealand, 2016.

[12] Free SFX Team, ‘Free SFX’, 2020. http://www.freesfx.co.uk (accessed Jul. 01, 2020).

[13] Qualtrics, *Qualtrics*. Provo, Utah, USA, 2005.

[14] R Core Team, *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria, 2019.

[15] D. Bates *et al.*, *lme4: Linear Mixed-Effects Models using ‘Eigen’ and S4*. 2020.

[16] G. Kidd, C. R. Mason, and F. J. Gallun, 'Combining energetic and informational masking for speech identification', *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 982-992, 2005.

[17] A. Braun, C. Llamas, D. Watt, P. French, and D. Robertson, ‘Sub-regional “other-accent” effects on lay listeners’ speaker identification abilities: A voice line-up study with speakers and listeners from the North East of England’, *The International Journal of Speech, Language and the Law*, vol. 25, no. 2, pp. 231–255, 2018, [Online]. Available: https://doi.org/ 10.1558/ijsll.37340.

[18] S. V. Stevenage, G. Clarke, and A. McNeill, ‘The “other-accent”effect in voice recognition’, *Journal of Cognitive Psychology*, vol. 24, no. 6, pp. 647–653, 2012, doi: 10.1080/20445911.2012.675321.

[19] H. Guner and C. Acarturk, ‘The use and acceptance of ICT by senior citizens: a comparison of technology acceptance model (TAM) for elderly and young adults’, *Universal Access in the Information Society*, vol. 19, pp. 311–330, 2020, [Online]. Available: https://doi.org/10.1007/s10209-018-0642-4.

[20] K. S. Helfer and R. L. Freyman, ‘Aging and speech-on-speech masking’, *Ear and hearing*, vol. 29, no. 1, pp. 87–98, 2008, [Online]. Available: https://doi.org/10.1097/AUD.0b013e31815d638b.

[21] N. Atkinson, ‘Variable factors affecting voice identification in forensic contexts’, PhD, University of York, 2015.

[22] P. Foulkes and A. Barron, ‘Telephone speaker recognition amongst members of a close social network’, *Forensic Linguistics*, vol. 7, no. 2, pp. 180–198, 2000.