How Frequency and Harmonic Profiling of a 'Voice' Can Inform Authentication of Deepfake Audio: An Efficiency Investigation

Emily L. Williams, Dr Karl O. Jones, Colin Robinson, Dr Sebastian Chandler-Crnigoj, Dr Helen Burrell and Dr Suzzanne McColl

Applied Forensic Technology Research Group
School of Engineering
Liverpool John Moores University
Liverpool, United Kingdom
E.L.Williams@2022.ljmu.ac.uk; K.O.Jones@ljmu.ac.uk; C.Robinson1@ljmu.ac.uk;
S.L.ChandlerCrnigoj@ljmu.ac.uk; H.Burrell@ljmu.ac.uk; S.M.Mccoll@ljmu.ac.uk

ABSTRACT

As life in the digital era becomes more complex, the capacity for criminal activity within the digital realm becomes even more widespread. More recently, the development of deepfake media generation powered by Artificial Intelligence pushes audio and video content into a realm of doubt, misinformation, or misrepresentation. The instances of deepfake videos are numerous, with some infamous cases ranging from manufactured graphic images of the musician Taylor Swift, through to the loss of $25 million dollars transferred after a faked video call. The problems of deepfake are becoming increasingly concerning for the general public when such material is submitted into evidence in a court case, especially a criminal trial. The current methods of authentication against such deepfake evidence threats are insufficient. When considering speech within audio forensics, there is sufficient 'individuality' in one's own voice to enable comparison for identification. In the case of authenticating audio for deepfake speech, it is possible to use this same comparative approach to identify rogue or incomparable harmonic and formant patterns within the speech. The presence of deepfake media within the realms of illegal activity demands appropriate legal enforcement, resulting in a requirement for robust detection methods. The work presented in this paper proposes a robust technique for identifying such AI-synthesized speech using a quantifiable method that proves to be justified within court proceedings. Furthermore, it presents the correlation between the harmonic content of human speech patterns and the AI-generated clones they produce. This paper details which spectrographic audio characteristics were found that may prove helpful towards authenticating speech for forensic purposes in the future. The results demonstrate that using specific frequency ranges to compare against a known audio sample of a person's speech, indicates the presence of deepfake media due to different harmonic structures.

KEYWORDS: *Artificial Intelligence, Digital Forensics, Speech Processing, Speech Analysis.*

1 INTRODUCTION

In the past decade, society has been driven forward by the rapid advancement of digital technology. This remarkable growth has raised several concerns across the world, with one notable example being the rise in fraudulent and deceptive practices such as deepfake creation (Armerding, 2017). A deepfake is a hyper-realistic, synthetically generated piece of media that can take many forms, often indistinguishable from real content. This poses a significant threat in several areas such as journalism, entertainment and cybersecurity as has been seen in recent years (Gregory, 2022) (Simmons, 2017). This manipulation of audio and video content, specifically the human voice raises questions regarding the authenticity of the media consumed by the public on a day-to-day basis. Furthermore, it brings into doubt the authenticity of evidentiary specimens within the criminal court and litigation (Europol, 2022.). Recent criminal cases of various natures showcase the growing use of fraudulent deepfake material, evidentiary by the 2019 heist that resulted in up to $35 million being stolen, by cloning the chief executive officer's voice (Stupp, 2019; Hertz & US Department of Justice, 2020). The use of deepfakes is incrementally becoming criminalized within UK law, especially with

https://doi.org/10.54389/HGBC7543

the Online Safety Bill, currently in its final stages of progression through the UK House of Lords (UK Parliament, 2023), however, there is still a significant lack of direct effective legislation to tackle this issue (Jones & Jones, 2022). In response to the growing threat, this research aims to address the issue of audio deep fakes within speech forensics, using a proposed method for frequency and harmonic profiling, doubling in use for speaker identity comparison in addition to authentication.

2 PROPOSED RESEARCH

2.1 Research Design

The work presented here uses a comparative approach to examine speech samples from individuals and generate AI-generated versions of their voices. The research explores the relationship between human speech patterns and the corresponding AI-generated voices. By employing a scripted template, the study identifies consistencies and inconsistencies within the variable results of imitated versions of each participant using a deep learning algorithm called Lyrebird, employed in the program known as Descript®. The resulting audio clips undergo spectrographic analysis, identifying the fundamental frequency at various salient points within the speech and further resulting harmonic structure within the voices. At these points, the fundamental frequency, minimum and maximum pitches are required, and a mean value is taken to summarize each point.

2.2 Participants

To accurately establish the relationship of harmonic structures between raw speech and the synthesized version, a specific group of volunteer English-speaking males between the ages of 18 – 60 were engaged to take part. Age is not a significant factor in this study, since previous research by Suzuki *et al*. (2022) has shown that specific frequency ranges (1-2kHz and 4-6kHz) remain unaffected by age and were crucial for investigating individuality. Therefore, particular attention was paid to these frequency ranges within this research, though comparisons were drawn between the original speech between each participant's sample. Participants were briefed on how to follow the script, using their normal speaking voice, without deliberately adding an accent or changing volume. The participants were also asked to speak in a somewhat monotone voice to gain an understanding of their natural fundamental vocal frequency and subsequently their harmonic structure as a result.

2.3 Data Collection

The recordings of speech were taken in a controlled environment to reduce background noise and interference. However, the recordings were treated to remove any remaining background interference to enable optimal signal analysis. Any post-recording noise reduction would reduce the viability of the voiceprints in the recordings hence this step was not taken to aid analysis (Harrisson *et al*, 2023).

Participants were instructed to read aloud a standardized text to ensure consistency across all samples in a similar set of research to Oshima *et al* (Oshima et al., 2016) however, this work was conducted in English as opposed to Japanese. These excerpts comprised a combination of the key 44 phonic sounds (Ellinas et al., 2023)(Miskin, 2020) and a selection of phoneme pangrams in addition to a recitation of the alphabet and counting between 1 through 20. This range was designed to allow for all possible standardized sounds that would be created within the English language. The phoneme pangrams were also designed to combine difficult-to-pronounce combinations of phonics, which in the context of this study, both aid possible unique pronunciation points (Ellinas et al., 2023) and further train the AI modelling software in how exactly the subjects speak in more difficult situations. In addition to these elements, the test includes a story-telling or constant speaking section consisting of an excerpt of the *Descript® Overdub* script that is used for subjects to provide at least 10 minutes of speech to create the AI model.

2.4 AI Model Creation

The collected speech samples were processed using Descript®, an AI-driven audio editing program. Descript® utilises an AI model known as Lyrebird which is deemed to be one of the best systems available for speech imitation. Once the relevant samples were uploaded to the Descript®/Lyrebird software, a period of 2-24 hours is allowed for the AI algorithm to run its course.

Once the AI model has been trained and the 'voice' is ready for use, a new set of phoneme pangrams were input to the text-to-speech (TTS) generator within Descript® to create an audio that can be used for comparison against the original 'raw' speech audio. This new set of phoneme pangrams forced the AI program to extract the sounds created within the 'raw' speech and rearrange them according to the text-to-speech input. This eliminated the possibility of a simple copy-and-paste of a large section of the original speech, therefore by breaking down the speech into its component phonic sounds, the harmonic structure is also broken in numerous places where the AI program needed to restructure according to its trained algorithm.

Figure 1. Typical Pipeline for TTS speech generation

Lyrebird is an algorithm based on the founders' own research. The system utilizes a Recurrent Neural Network (RNN) (Mehri et al., 2017) to provide an end-to-end unconditional audio synthesis model. Once the result is returned, a usable model is presented. A series of imitation scripts were used to extract specific sounds, especially vowels, (Oshima et al., 2016)(Zhang et al., 2019) which acted as analysis points within the audio and speech. The original recordings and the deepfake versions were then loaded into a formant analysis program that extracted the values of the dominant harmonics within the comparable audio clips.

2.5 Harmonic Structure Extraction

To extract the harmonic structure of each speech sample, software called 'Praat' was selected. Praat is a system that works using a Linear Predictive Coefficient (LPC) weighted spectral matching to identify speech within a signal. It can extract information using Fast Fourier Transport (FFT) and pick out the formant values from the LPC calculations within the algorithm (Wood, 2020). The Fast Fourier Transport enables a spectrographic view of the signal, which corresponds to the formant values identified by the LPC.

Praat has the capability to adjust the sensitivity of formant detection, with the application of a set number of formants to enable consistent resolution allowing continuity of handling across all audio samples whilst taking into account varying speech envelopes, namely Attack, Decay, Sustain & Release (ADSR). The program originating from the University of Amsterdam has previously been used within clinical and research settings making it a viable program to use for acoustic speech analysis within this research.

Within the recordings, certain sections of the speech were selected for specific comparisons. Several studies investigated these comparisons, and their conclusions indicate that the most salient points for vocal comparison were vowels present within the language (Zjalic, 2021; Suzuki et al., 2022; Kaiser & Bořil, 2018) . Given the set structure provided to the participants for the speech recording, it is possible to pinpoint the exact instances of vowel pronunciation. These vowel speech 'nodes' were edited to extract from the surrounding material and to compare against the AI model version of these 'nodes' within the Praat software. The resulting spectrographic representations and frequency band 'formant' values were compared. Since there were multiple instances of these vowel 'nodes' within the speech, a number of these were extracted and their formant values were identified. Within this, an average value for each vowel type was calculated and compared directly with the similar value for the AI speech. This indicated the error of the harmonic speech at the vowel node points. Furthermore, it

revealed any structural differences between the formant ‘bands’ that appear in the values provided by Praat.

2.6 Data Analysis

Within the works of Suzuki et al (Suzuki, Ishimaru, Toyoshima, & Okada, 2022), it is argued that the main ‘individuality’ frequency markers were within the 1-2kHz and 4-6kHz ranges, however as the main part of human speech’s frequency content occurs within the 100-4kHz range (Yost & Yost, 2000), any frequencies above 4kHz were considered ‘secondary’ harmonics within this research as they are not direct contributions of the fundamental frequency sequence and structure within the speech and thus analysed separately from the frequency structure below 4kHz.

Statistical measures, such as fundamental frequency (F_0) variations, harmonic-to-noise ratios and format structures & values were calculated and compared objectively. First, the ‘raw’ original samples were be compared with one another, in addition to the ‘raw’ original speech samples being compared against their resulting deepfake version. This provided an overarching view of the differing values as a result of AI processing within the frequency structure. It may be possible to analyse a unique structure that is formed within the deepfake speech files only which may prove to be a marker of such media itself within the future, though this would require further research across numerous different AI algorithms and programs.

The qualitative element of this research was to analyse the patterns of formants (harmonics) themselves within the spectrographic time/energy domain. Once a quantitative value for each formant is extracted, a visual manifest of the patterns within the same domain was compared to allow for tonal differences between the two data sets.

2.7 Ethical Considerations

Since this research is attempting to create a deepfake version of someone’s voice, considering the implications of the misuse of a voice clone, there is utmost importance placed on the security and anonymity of the voice samples provided. To protect the individuals taking part in the research, their name is only requested to take a vocal sample of permission given to use the AI program selected to conduct the research. Participants were always allowed the right to withdraw at any time and their files will be deleted upon request. Extra care is taken to ensure the protection of the participant's rights and privacy. Participants’ data is stored securely, with restricted access.

2.8 Limitations of approach

There were several limitations associated with this research. Firstly, there may be differences within recording environments, despite efforts to make them correspond, which impacts the overall result of the recordings and ultimately, the quality of the deepfake audio file within the AI algorithm.

The nuances of human speech also impact the results of this research. As has been shown in previous research, and the basis upon which this research is founded, each human voice is totally individual. However, there were some instances where voice imitations conducted by a human (Voice impression artists) can sound extremely like someone else (Kitamura, 2008).

Other human factors also impact the quality of the ‘raw’ speech samples collected, including illness, surgery, environment, and other such varying factors. Whilst elements such as age and ethnicity have already been addressed in this paper, there is currently no viable way to observe whether a participant is ill, and their voice has been affected. Considering that illness does have an impact on vocal quality, there may arise an issue whereby the developed technique of analysis cannot handle the impaired human voice when comparing human-to-human speech samples. At this point, there would need to be further investigation into the limits of the comparison technique when the participant is suffering from a long-term illness. Since there is significant research on the possibility of ‘Biomarkers’ within speech that may indicate the presence of certain diseases and illnesses (Ramanarayanan, Lammert, Rowe, Quatieri, & Green, 2022), this may, in turn, allow for further identification if there were any significant markers identified within a spectrographic approach to ‘biomarking’.

In addition, the use of scripted speech may hinder some of the natural nuances of free speech (Stevenage, Tomlin, Neil, & Symons, 2021). However, Since Human speech is not inherently planned

in the speaker's mind (Hays, 2014), the use of scripted speech alleviated any anxiety in the participant's mind regarding what they said in addition to providing a structure of sounds to be used within the analysis of this research. The difference in the style of speech (Scripted vs. free) would hinder any analysis during the comparison of raw speech vs. raw speech recordings, or even live speech. However, since this research is investigating the use of deepfake media, there is no secondary human speaker per se. The secondary speaker is essentially a computer-generated voice and thus requires a script to generate sound, which in turn, warrants a scripted recording of speech to aid a more direct comparison between the human vs. deepfake discrimination.

Within this research, a total of 200 'node' points (identified frequencies) were identified per participant, allowing for 20 for each eventuality of vowel sound presented within the AI-generated speech and the natural (referred to as Raw within this research).

Similar to the research in (Yoo, Lim, & Yook, 2015), voice detection is based around the identification of formants within the file. These formants indicate the presence of the harmonic structure within the voice. The research in (Yoo et al., 2015) analyses 3 values of formants for vowel sounds to detect speech, however, in this research, it is prudent to extend that number to a maximum of 5 values to enable a wider profile of harmonics to be identified.

3 RESULTS

The initial 'raw' samples obtained from each participant were clear and contained little 'noise' however some samples were more 'noisy' than others. All participants spoke with a 'gently monotone' voice, where there was no apparent loss of natural vocal expression. There was no noise reduction required on these samples before being sent for processing within the Descript® program. All files were recorded at a minimum of 44.1kHz sample rate and 16-bit depth. This was specifically chosen to cater for the standard audio that a criminal investigation may receive. These specifications have been the accepted 'standard' of audio since the inception and success of the CD in the music industry. This quality of recording allows for Nyquist's theorem to be applied to the entire audible range of sound whilst allowing headroom in addition (Yu, 2019).

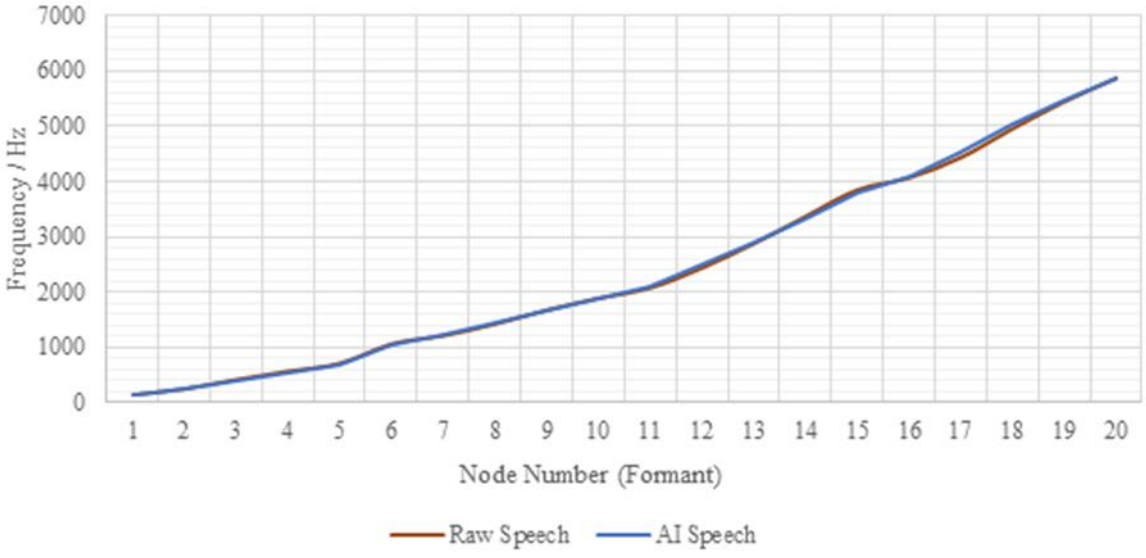

Figure 2. Average speech 'profile' presenting a structure of frequency content.

The values for the 'nodes' or F points, in the results of this research, were taken using the Praat software, where the values were extracted using the formant detection algorithm. These values were corroborated against the spectrographic values of the harmonics to ensure that the signals were indeed present within the audio file, ensuring that they were not in the presence of extraneous sound.

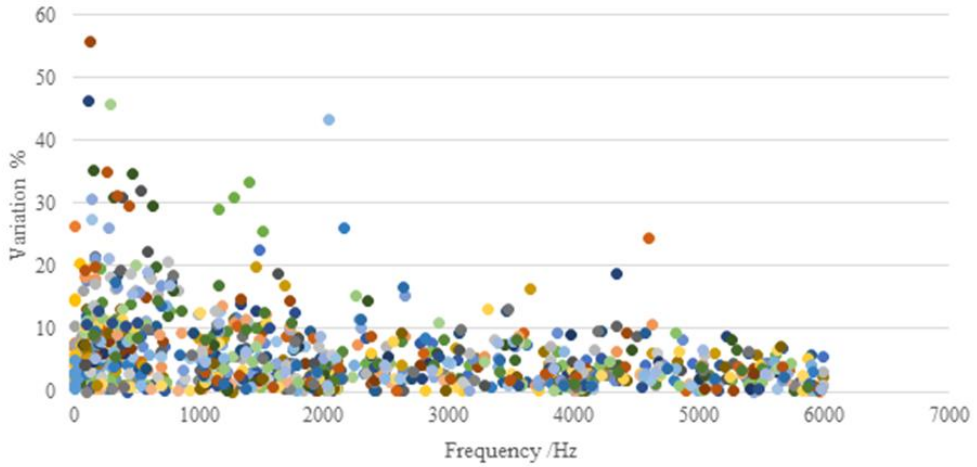

Figure 3. Variation between AI and Raw speech data sets presented as % difference.

Initial results were collected for the first participant to test the theory of how the Descript® Lyrebird program would process the samples uploaded. These results were only presented for the first 5 sample points (nodes) within the profile generated.

The raw data presented in Appendix 1 includes the values of each sample point within both the raw speech and the AI-generated speech. The averages of these values have been used to present Figure 2 where the average profiles of both the AI and raw speech results were compared. This 'structure' represents the efficacy of the AI model in addition to the proposed method for the authentication of AI. Where the average results were similar in Figure 2, Figure 3 presents a more accurate view of the differences found between both speech data sets. Figure 3 shows that there is some variation between the results and that the AI model did not create a 100% accurate reproduction where the imitation is faultless as shown in Table 1.

Table 1. The frequency content of two vowel samples from Participant #1 vs. AI generated version Where /x/ indicates the vowel sound and 'x' indicates the word derived from.

Name	**F0 *Hz***	**F1 *Hz***	**F2 *Hz***	**F3 *Hz***	**F4 *Hz***
/a/ 'sat' raw	92	166	447	620	732
/a/ 'sat' AI	97	156	401	559	679
/e/ 'egg' raw	100.5	171	388	532	692
/e/ 'egg' AI	93	158	408	540	675

In total, 10 participants were willing to engage with this research and provide a sample of their voices, which in turn produced 20 profiles to be compared against each other, both inter-original and inter-generated. These participants were all male adults of English-speaking origin however, there was some variation of accents including English regional and Scottish. None of the participants attempted impersonation or vocal imitation, including altering of their accents. All participants used their natural voice in a calm and consistent manner.

Assessing the potential for the use of harmonic content analysis for the purposes of identity verification, the samples taken from each participant were compared with one another without the samples of the AI deepfake dataset. It was apparent that speaker 8 had the highest pitch voice of all participants, and whose harmonic values were generally higher than the others. Although this profile deviates from the others in the results, it was still included in the mean profile being analysed in comparison to the AI-generated results. Therefore, it is important to consider its impact on the overall

data. Figure 2 shows the variation that was identified within each frequency sampled from the raw speech. It represents the percentage difference between the raw vs. AI at each nodal point. Figure 2 shows that there is consistent variation between 2-6kHz, but there is a large array of variation in the lower frequency ranges. It demonstrates that there is more predictability and consistency in the higher frequencies than in the low ones. Where AI is concerned, it is more likely to have a more consistent difference in frequencies above 2kHz than below.

4 DISCUSSION

The possible implications of this method of voice comparison/ analysis may be constricted due to the low possibility of high-quality, low-noise samples of audio being submitted as evidence. Whilst this may be treatable to a small extent with the use of noise removal, there is in turn a high likelihood that this would impact the results of the analysis by removing instances of frequency content that were integral to the structure of one's voice. Realistically, this proposed technique for analysis could be used in situations where higher quality audio has been supplied, for example, recording via covert equipment, and audio from digital media, especially social media. This method may however be helpful in certain situations where restricted frequency content is available such as telephone communication, Voice Over Internet Protocol (VOIP), recorded voice messages (What's App, Messenger, etc.) and legacy media. Since most media in this context is focused specifically on the range of speech, it is possible to reconstruct the frequency content, considering the application applied originally. Even if the reconstruction is not possible, there is a high likelihood that the vocal profile could be extracted from the audio file to gain an insight into the person (or machine) speaking.

The integration of this process within the current workings of law enforcement would not be difficult, but another step within the task of authentication of evidentiary origin, ensuring the 'chain of evidence' has been upheld, as required by law (British Standards Institute, 2017). Within the process of authentication, the origin of evidence must be called into question. In utilizing this process in the authentication of speech, deepfake speech is likely identified. It can be recommended that this proposed technique for identification should be called upon when suspect material is identified.

The applications of this research are also not entirely restricted to the criminal use of AI and deepfake. There are many other instances where deepfakes have been used in non-criminal areas, such as entertainment, Film, TV, Radio, and the Music industry for positive reasons. In conjunction with this proposed method of authentication, there are also proposed methods for ensuring that deepfake files can be easily identified by anyone. One such suggestion is the automatic inclusion of an inaudible signal within the file which the specific constant frequency in the file corresponds to a pre-determined deepfake generator. There are several issues with this. Firstly, audio can be edited. If one knows of the existence of these signatures, then a simple EQ could remove them. Secondly, how is it enforceable? There would need to be legislation in place to ensure this takes place, but even in this circumstance, there will likely be software generated that can subvert the legalities and not include the signatures. If a recording were checked by a governing body and was found to not have the signature as required, how is it then proven that this does in fact come from the software that is being tested? Thirdly, there would need to be some method of organization to arrange the prescribed frequencies for each software. There is the constant risk that by the pure ability of AI to learn every time it is used, owing to the deep learning ability, AI would inevitably get to a point where it is almost indiscernible from real speech. Purely by this matter, measures should be put in place as previously mentioned in this research.

There also exists the potential for criminals to utilize other systems that don't necessarily clone a voice but could use voice morphing, and other methods of voice impersonation, especially speech-to-speech voice cloning. Unless this is recorded at the time of being used, there is no way to currently identify that this occurring in real-time, for example, if someone is disguising their voice using speech-to-speech voice cloning and accessing a bank account over the phone.

Since speech is a moving subject, no one speech profile fits one their whole life as their voice will change under several factors. This is a concern when using the proposed approach for inter-person clarification, since both samples (exemplar and evidentiary), need to be from the same time period and not subject to any personal changes such as major illness or medical procedures etc. The same could be said from a deepfake authentication perspective. There is a likelihood of identifying when such a deepfake was made within the metadata of the file itself, which would then enable to sampling of speech

audio from the same time period. In the case that the deepfake was made with criminal intent, to impersonate their victim, it would be possible to consider what public access samples of the victim's speech are available and consider those for the authentication process.

5 CONCLUSION

The objective of this study was to assess the correlation between the harmonic content of human speech patterns and the AI-generated clones they produce. Through thorough data collection, the creation of AI models, and the identification and extraction of harmonic structures, valuable findings were obtained. These findings uncovered several characteristics that may prove helpful in authenticating speech for forensic purposes in the future.

Within the AI model creation, it was highlighted just how convincingly the Descript® software cloned and imitated the human voice. Utilizing Praat, we investigated the underpinning structures of both human speech and AI-generated voices, uncovering intricate patterns and individualistic points of interest, shedding light on areas that could be used as an indication point during identity comparison and deepfake authentication based on these uniqueness factors. Both human speech and AI have a pattern within their structure which may be telling of their existence in the future.

When exploring the correlation between the data sets, both quantitative and qualitative data were considered, with the quantitative data providing an explicit value for the harmonic positions within the speech, the qualitative approach identified several visual factors that could also play a significant role in identifying the whereabouts of deepfake media within a suspect file.

The subject of ethics was considered throughout, considering the risks involved with voice cloning and reproduction. The anonymity of the participants has been preserved and any link between the participants and their speech profile has been severed. Access to the information within these recordings has only been accessed by the authors of this research. This ethical responsibility must be highlighted in association with the fraud risk in the event of the recordings and AI model data being obtained by a third party.

In conclusion, this research contributes to the growing area of deepfake authentication and identification. Though there is a significant threat of AI technology developing to such a point where it is entirely indistinguishable from authentic original speech, this research aims to highlight the importance of the awareness of such a threat. Preventative measures must be put in place to ensure the future safety of one's vocal identity, whether it be in a professional or personal capacity, both are of utmost importance to protect against the threat of imitation and exploitation.

The proposed technique of extracting the harmonic structure information and comparing it to a known sample at specific points within the speech would significantly increase the chances of identifying deepfake material within the initial steps of investigation. Given that there is proven variation of the deepfake material being 3-10% different to the original raw speech, particularly within the 1-2kHz and 4-6kHz ranges, there can be identification of deepfake speech on this basis. It is recommended however that further research take place into interpreting a wider data set and presenting more diverse structures according to the context of the speaker. Further research should also take place into the specific and measurable changes that illness, pitch shifting, frequency redaction and other aspects of manipulation would further hinder the identification of deepfake materials within evidence.

REFERENCES

Armerding, T. (2017, May 16). *Vocal theft on the horizon : Using your voice for authentication is about to get more risky, thanks to voice-spoofing technology*. https://www.csoonline.com/article/561705/vocal-theft-on-the-horizon.html

British Standards Institute. (2017). BS EN ISO / IEC 17025 : 2017 BSI Standards Publication General requirements for the competence of testing and calibration laboratories, 2017(June), 1–47.

Ellinas, N., Christidou, M., Vioni, A., Sung, J. S., Chalamandaris, A., Tsiakoulis, P., & Mastorocostas, P. (2023). Controllable speech synthesis by learning discrete phoneme-level prosodic representations. *Speech Communication*, 146(November 2022), 22–31. Retrieved from https://doi.org/10.1016/j.specom.2022.11.006

Europol. (2022). *Facing reality? : law enforcement and the challenge of deepfakes : an observatory report from the Europol innovation lab.* https://doi.org/10.2813/08370

Forensic Science Regulator (2020). Codes of Practice and Conduct FSR-C-134, (1), 1–15.

Gregory, S. (2022). Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism*, 23(3), 708–729. Retrieved from https://doi.org/10.1177/14648849211060644

Harrisson, O., Reed-jones, J. T., Morrison, K., Robinson, C., & Jones, K. (2023). The Effect of Noise Reduction upon Voiceprint Integrity. *International Conference on Intelligent Systems and New Applications*, 211–217

Hays, B. (2014). You don't really know what you're talking about, scientists say. Retrieved from https://www.upi.com/Science_News/2014/05/02/You-dont-really-know-what-youre-talking-about-scientists-say/9331399051132/#:~:text=But some cognitive scientists have,they%27ve already said it.

Hertz, R. G., & US Department Of Justice. (2020). *Reference: DOJ Ref. # CRM-182-77215*. Columbia District.

Jones, K. O., & Jones, B. S. (2022). *HOW ROBUST IS THE UNITED KINGDOM JUSTICE SYSTEM AGAINST THE ADVANCE OF DEEPFAKE AUDIO AND VIDEO?* (Vol. 55606). Retrieved from http://insight.cumbria.ac.uk/id/eprint/6675

Kaiser, J., & Bořil, T. (2018). Impact of the GSM AMR Codec on Automatic Vowel Formant Measurement in Praat and VoiceSauce. *2018 41st International Conference on Telecommunications and Signal Processing, TSP 2018*, 1–4. Retrieved from https://doi.org/10.1109/TSP.2018.8441185

Kitamura, T. (2008). Acoustic analysis of imitated voice produced by a professional impersonator. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 813–816. Retrieved from https://doi.org/10.21437/interspeech.2008-248

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., … Bengio, Y. (2017). Samplernn: An unconditional end-to-end neural audio generation model. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–11.

Miskin, R. (2020). *Read Write Inc. Phonics: Reading Leader Handbook* (2nd ed.). Oxford University Press.

Oshima, Y., Takamichi, S., Toda, T., Neubig, G., Sakti, S., & Nakamura, S. (2016). Non-native text-to-speech preserving speaker individuality based on partial correction of prosodic and phonetic

characteristics. *IEICE Transactions on Information and Systems*, E99D(12), 3132–3139. Retrieved from https://doi.org/10.1587/transinf.2016EDP7231

Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F., & Green, J. R. (2022). Speech as a Biomarker: Opportunities, Interpretability, and Challenges. *Perspectives of the ASHA Special Interest Groups*, 7(1), 276–283. Retrieved from https://doi.org/10.1044/2021_persp-21-00174

Simmons, D. (2017). BBC fools HSBC voice recognition security system. Retrieved 11 July 2023, from https://www.bbc.co.uk/news/technology-39965545

Stevenage, S. V., Tomlin, R., Neil, G. J., & Symons, A. E. (2021). May I Speak Freely? The Difficulty in Vocal Identity Processing Across Free and Scripted Speech. *Journal of Nonverbal Behavior*, 45(1), 149–163. Retrieved from https://doi.org/10.1007/s10919-020-00348-w

Stupp, C. (2019). Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Retrieved 7 September 2023, from https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Suzuki, N., Ishimaru, M., Toyoshima, I., & Okada, Y. (2022). Identifying Voice Individuality Unaffected by Age-Related Voice Changes during Adolescence. *Sensors*, 22(4), 1–16. Retrieved from https://doi.org/10.3390/s22041542

UK Parliament. (2023). Online Safety Bill 2023.

Wood, S. (2020). Praat for Beginners: Formant tracking in the Sound editor. Retrieved 12 September 2023, from https://swphonetics.com/praat/snded/formtrack/#:~:text=Praat for Beginners%3A,frequencies on the spectrogram itself.

Yoo, I. C., Lim, H., & Yook, D. (2015). Formant-based robust voice activity detection. *IEEE Transactions on Audio, Speech and Language Processing*, 23(12), 2238–2245. Retrieved from https://doi.org/10.1109/TASLP.2015.2476762

Yost, W. A., & Yost, W. A. (2000). *FUNDAMENTALS OF HEARING* (Vol. 3). Academic Press.

Yu, R. Q. (2019). Sampling | theory. *Encyclopedia of Analytical Science*, 143–149. Retrieved from https://doi.org/10.1016/B978-0-12-409547-2.14109-5

Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R. J., … Ramabhadran, B. (2019). Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-Septe, 2080–2084. Retrieved from https://doi.org/10.21437/Interspeech.2019-2668

Zjalic, J. (2021). *Digital Audio Forensics Fundamentals - From Capture to courtroom*. Routledge.