applied sciences

MDPI

Article

Evaluation of the Human Capacity to Detect Spanish Deepfake Audios with a Paraguayan Accent

María Vianella Giménez Ramos [1,*], Juan Pinto-Ríos [2], Pastor Pérez-Estigarribia [1,3] and Enrique Dávalos [1,*]

[1] Facultad Politécnica, Universidad Nacional de Asunción, San Lorenzo 111421, Paraguay; pastor.perez@sudamericana.edu.py
[2] Escuela de Ingeniería Eléctrica, Pontificia Universidad Católica de Valparaíso, Av. Brasil 2950, Valparaíso 2340000, Chile; juan.pinto.r@pucv.cl
[3] Facultad de Ciencias de la Salud, Universidad Sudamericana, Pedro Juan Caballero 1206, Paraguay
* Correspondence: vianellagimenez@fpuna.edu.py (M.V.G.R.); edavalos@pol.una.py (E.D.)

Abstract

Deepfakes, synthetic multimedia files generated by artificial intelligence, are drastically undermining digital credibility. Their ability to manipulate our perception of reality has created a new and complex battleground for disinformation, posing a critical threat to non-English-speaking audio with distinctive accents. Consequently, the objective of this study is to determine the human capacity to detect deepfake audio in Spanish with a Paraguayan accent through an experiment conducted with an Android application called ReFake (developed specifically for this research). In this experiment, 450 participants, aged 16–72, evaluated 10 audio samples of up to 15 s each, classifying them as authentic (belonging to Paraguayan journalists) or fake (generated with ElevenLabs). The findings suggests that human ear is more accurate than artificial intelligence (AI) at detecting vocal 'naturalness'. This ability is influenced by generational age and educational level, with younger people and those with postgraduate degrees demonstrating greater performance. Conversely, gender and nationality do not influence detection, although the high prosodic quality of deepfakes still leads to errors in human judgment. Given these results, it is crucial to adapt and develop new strategies for a secure and resilient online ecosystem.

Keywords: artificial intelligence; audios; cybersecurity; deepfakes; human detection

1. Introduction

The past decade has seen an unprecedented proliferation of synthetic media generated, in part or in whole, by generative artificial intelligence technologies. Specifically, digitally altered or completely fake content, known as "deepfakes" (a portmanteau of deep learning and fake), encompasses multimedia files (videos, images, and audio) that convincingly depict events, statements, or actions that never occurred. This hyper-realistic falsification capability has captured global attention, raising serious concerns about its potential for misuse and its social, political, and economic impact [1–7].

However, while deepfake technology has a wide range of potential applications that are still unexplored [4,8], promising a radical transformation in key sectors such as media, entertainment, education, and medicine, its widespread use raises concerns about the erosion of public trust in the veracity of digital multimedia content. This could catalyze a scenario of epistemological collapse, where misinformation is perceived as authentic and truthful content is denounced as fake. This increased realism of deepfakes introduces

Check for updates

Academic Editors: Hyokyung Bahn and Douglas O'Shaughnessy

Received: 17 November 2025
Revised: 30 January 2026
Accepted: 11 February 2026
Published: 14 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

emerging risks whose nature and scope conceptually transcend specific domains and contexts, necessitating a threat mapping that is still in its infancy [2–5,9,10].

Given this adverse outlook, the implementation of educational initiatives, regulatory frameworks, technical solutions, and the consideration of ethical implications represent potential and necessary avenues for addressing the challenge of deepfakes. Specifically in the area of detecting these synthetic media, significant limitations are observed in the field of research on audio deepfakes, particularly for those generated in languages other than English or featuring distinctive accents [11–13]. This disparity is due to the fact that the majority of studies (approximately 90%) have focused on developing machine learning algorithms for classifying images and videos as authentic or fake [8,14–16].

Continuing with the aforementioned limitations, previous research has consistently identified that current detection tools suffer from deficiencies in their generalization and robustness [17–19]. This is primarily attributed to significant and inappropriate biases present in the datasets used during the training process. Specifically, insufficient attribute quality, diversity, and representativeness in training and testing datasets has been shown to lead to disproportionately low performance and biases against specific demographic groups. These underrepresented populations notably include women, racial and ethnic minorities (particularly people of color), minors, and non-English speakers [8,20]. The lack of adequate representativeness in training data exacerbates the issue of bias, negatively impacting the fairness and applicability of these tools in diverse population contexts [21].

Considering that, to date, the current state of research raises significant questions and that technical solutions are not sufficiently developed to effectively address the threat of deepfakes [5,22–24], the primary objective of this study was to determine the human capacity to detect deepfake audio in Paraguayan-accented Spanish. Based on the findings, we seek to propose comprehensive strategies (technological and regulatory approaches) aimed at avoiding or reducing risks and promoting the ethical use of these technologies. Special attention was paid to contextualizing gaps in detection performance, identifying how certain personal characteristics of participants and audios could negatively influence their ability to differentiate authentic audio from deepfakes.

Given the current critical scenario of synthetic representations, which pose serious threats to global security, a series of questions and queries have been formulated that align directly with the essence of the general purpose of this work. This attempt aims to provide a coherent response to the factors, situations, or events that could impact, positively or negatively, the performance of human classification with respect to deepfake audio. Within this context, the project will seek to answer the following research questions:

- What demographic factors (such as gender, generation, education level, or nationality) influence people's ability to detect fake audio generated by artificial intelligence?
- How does the gender of the speaker and the gender of the listener influence the reliability of authenticity judgments, and is there a significant interaction between these two variables?
- How effective are human evaluators compared to artificial intelligence models in discriminating between authentic and synthetic audio?

Based on the above, the remainder of the article is organized as follows: Section 1 presents the context of the study, focusing on the current problems and limitations in deepfake detection. Section 2 describes the proposed research method, covering dataset acquisition, test design, and data analysis strategy. Section 3 presents the results in detail. Section 4 analyzes the significance of the key findings, the limitations of the study, and proposes future research directions. Finally, Section 5 summarizes the main conclusions.

2. Materials and Methods

2.1. Dataset

Given the lack of a specialized public dataset containing authentic and synthetic audio (deepfakes) in Spanish, a dedicated dataset was created. This new resource was developed based on the following technical and methodological specifications:

2.1.1. Authentic Audios

The collection of authentic audio samples was conducted in October 2023 at the facilities of a Paraguayan television network, using a Sony ICD-PX240 (Sony Corporation, Tokyo, Japan) digital voice recorder. MP3 files ranging in length from 1 to 88 s were generated, reciting in Spanish and a bit of Guarani, iconic phrases, emotions, and thoughts of each person. As such, the dataset was acquired in an uncontrolled environment (outside of a recording studio) through semi-structured and spontaneous interviews. The participants were 15 professionals from the Paraguayan journalism industry (6 men and 9 women), aged between 30 and 75 years.

Under the premise of normalizing and standardizing these multimedia files, recorded audio clips that exceeded the 15-s threshold were manually edited. The open-source tool Audacity 3.4.2 was used to do this, adjusting each clip to this predetermined maximum length. Consequently, a total of 45 audio clips were created for testing and analysis (three authentic clips for each person who participated in the recording sessions).

2.1.2. Deepfake Audios

The ElevenLabs Starter version commercial software was used to generate the synthetic multimedia content, employing the Instant Voice Cloning method. The process was carried out using the Eleven Multilingual v2 model, with a single 15-s sample of authentic audio from each of the 15 participants as a reference.

In order to ensure the reproducibility of the study, the synthesis was performed under a standardized technical configuration of the voice controls (Voice Settings), using the platform's default values: Stability at 50%, Clarity + Similarity Enhancement at 75%, Style Exaggeration at 0% and with the Speaker Boost option enabled.

The synthesis process was performed without prior training (zero-shot), using grammatically correct text fragments tailored to each subject's personality and linguistic style. After generation, the files were exported in MP3 format. Finally, a manual post-processing stage was carried out in Audacity 3.4.2 software to remove residual noise from the synthesis tool (clicks or artifacts at the audio edges) and normalize the amplitude levels, thus ensuring sonic parity with the authentic samples during testing.

2.2. Procedure

2.2.1. Design and Sampling

In order to carry out observation and data collection tasks for subsequent statistical analysis, a native Java-based mobile application for Android called ReFake (used in version 1.1.4) was developed. Its backend architecture relies on the database and user authentication services of Google's cloud platform, Firebase. Participants registered in this application by providing certain personally identifiable information about themselves, such as their first name, last name, date of birth, gender, nationality, educational level, and email address.

Once the initial steps were taken, each participant was presented with 10 different, randomly selected audio clips (labeled as "authentic" and "fake"), in which the test subjects listened to an equal number of real and synthesized audio clips (5 of each type), but were not informed about the actual proportion of stimuli. More specifically, the key aspect was a binary-response question to classify a stimulus as benign or not.

Expanding on what was described in the previous paragraph, the study was designed to strike a good balance between favorable accuracy and minimal risk of cognitive overload, limiting the sample size to only 10 audio samples per participant. Furthermore, for each classification, the confidence level of the responses was required on a 5-point Likert scale (from none to complete confidence). Finally, participants were given the option to answer a free-text question to explain the reasoning behind their choice, as detailed in Figure 1.

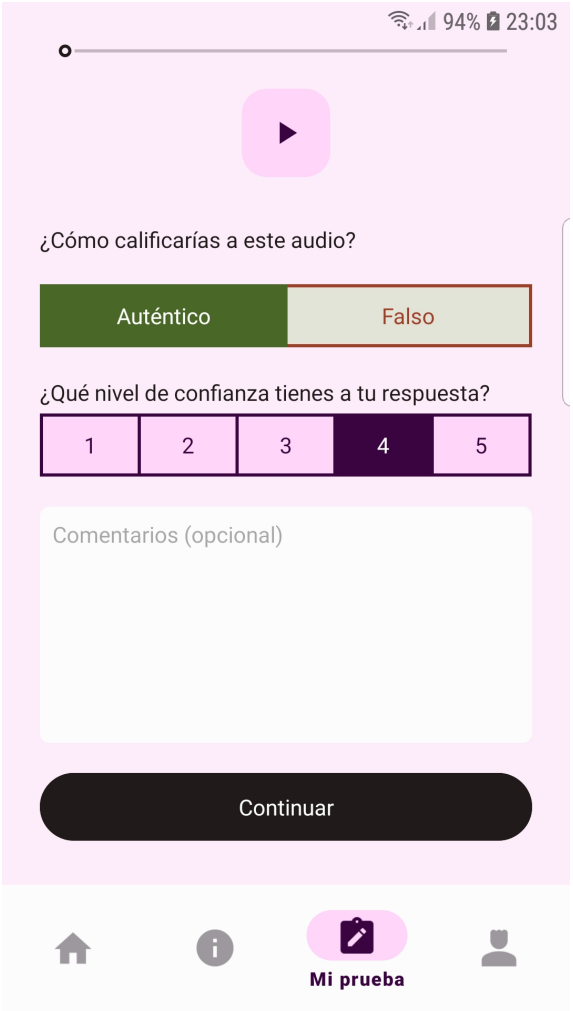


Figure 1. User interface of the ReFake mobile application during testing. The Spanish terms shown are: *¿Cómo calificarías a este audio?* (How would you rate this audio?), *Auténtico* (Authentic), *Falso* (Fake), *¿Qué nivel de confianza tienes a tu respuesta?* (What is your confidence level in your response?), *Comentarios (opcional)* (Comments (optional)), *Continuar* (Continue), and *Mi prueba* (My test).

It's worth noting that users completed a single round of testing in which the sequence of 10 randomly displayed audio samples originates from a general data set comprising 90 samples (45 authentic and 45 fake). Finally, once the entire test set is completed, the app displays a summary of the score achieved along with a history of all correct and incorrect answers, detailing the categorized audio and whether it corresponds to authentic or fake.

2.2.2. Participants

Between 11 January 2024, and 4 January 2026, observational tests were conducted with a group of 450 volunteers. This sample consisted of people close to the authors and passersby randomly selected from public places, as well as members of online communities. Participants were not offered any monetary incentive or compensation for taking part in the research.

Although the vast majority of users have been recruited through information distributed in WhatsApp groups, a portion of them come from educational centers in various cities in Paraguay. Continuing these same lines, the general sample resulting is characterized by a multiplicity of demographic and academic parameters, as broken down in Table 1.

Table 1. Summary statistics for the dataset.

		Participants	Percentage
Gender	Female	238	52.9%
	Male	212	47.1%
Generation	Baby Boomers	97	21.6%
	Generation X	116	25.8%
	Millennials	123	27.3%
	Generation Z	114	25.3%
Education	Basic school education	110	24.4%
	Secondary education	132	31.1%
	University education	125	27.8%
	Postgraduate	83	16.7%
Nationality	Foreigner	130	28.9%
	Paraguayan	320	71.1%

2.3. Data Analysis

2.3.1. Work Environment

The entire analysis was developed entirely using the free version of the Google Collaboratory (Colab) cloud platform. The computations were executed using the Python 3 programming language.

2.3.2. Preprocessing

The raw data, extracted from observational tests in a semi-structured format (JSON), consisted of two elements: previously anonymized personal information (excluding direct identifiers such as first name, last name, and email address) and the responses recorded in the survey. From this, the first step was to perform the initial conversion from JSON to a tabular structure (Excel format). A debugging process was then carried out to identify and eliminate atypical, inconsistent, duplicate, or incomplete data (missing values). The audio recordings were divided into three ranges: short, medium, and long, with cutoff points of 5, 10, and 15 s. This resulted in a unified data structure, optimal for multidimensional (parallel and cross-sectional) comparative analysis between treatment groups.

2.3.3. Evaluation Metrics

A binary classification (positive and negative) was defined in Figure 2, linking the correct answers and errors of the subjects with the authentic labels in the data set. Thus, the elements of the main diagonal correspond to those correctly identified (green boxes), and those that are outside (red boxes) do not.

Table 2 presents the six performance metrics focused on in this study: accuracy, precision, specificity, sensitivity (also known as true positive rate, TPR), false positive rate (FPR) and F1 Score. All of these indicators take values within the range of 0 and 1, where 1 represents optimal performance (perfect classification). A score of 0.5 reflects performance attributed to luck or random guessing. For a result to be significant, the test must exhibit a score substantially higher than 0.5. A result below 0.5 suggests that the test is providing evidence against the hypothesis [25].

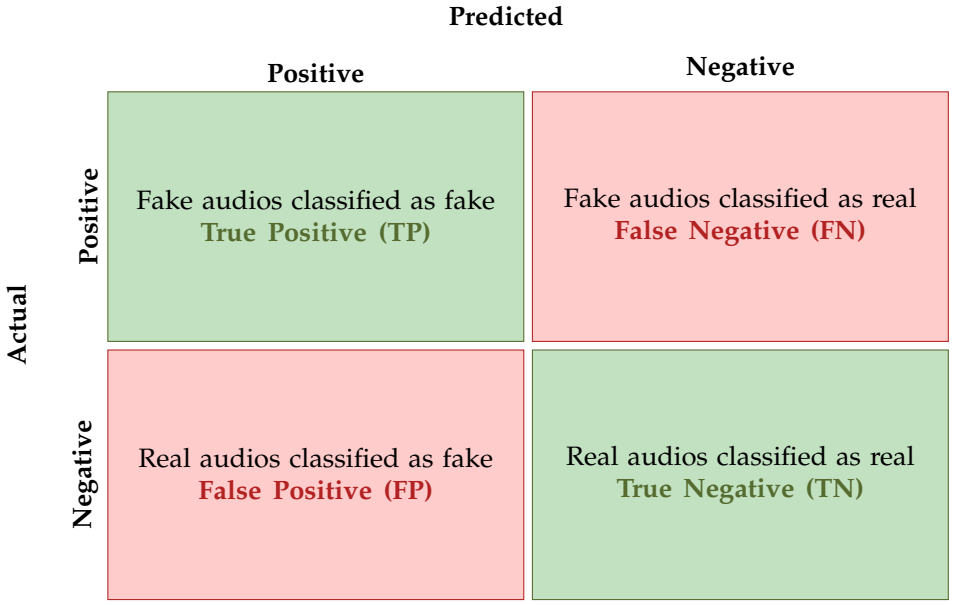


Figure 2. Binary confusion matrix.

Table 2. Performance metrics based on the confusion matrix [25].

Metric	Equation	Evaluation focus
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	Proportion of total correct predictions (overall performance)
Precision	$\frac{TP}{TP+FP}$	Measures the quality of positive results, is crucial to minimizing FP
Specificity	$\frac{TN}{TN+FP}$	Effectiveness of a diagnostic model or test to identify negative instances
TPR	$\frac{TP}{TP+FN}$	Ability to find all positive samples (if a high TPR is prioritized, the FN is sought to be avoided at all costs)
FPR	$\frac{FP}{FP+TN}$	Risk of generating a false alarm, a low FPR indicates good specificity
F1 Score	$2 \times \frac{precision \times TPR}{precision + TPR}$	Finds the balance between precision and TPR, useful in class imbalance

2.3.4. Methodology of Statistical Analysis

The initial phase of the analysis consisted of summarizing the responses from the dataset by participant and audio. It should be noted that these responses were represented in the confusion matrix in Figure 2 and coded using a binary, mutually exclusive scheme (one-hot encoding). From this, each of the six metrics specified in Table 2 was calculated individually. Subsequently, these results were subjected to a statistical significance analysis to evaluate the impact of demographic variables on performance. The procedure began with Levene's test to verify the homoscedasticity of the data, a criterion that determined the selection of omnibus tests. For comparisons between two groups, Student's t-test (or Welch's t-test for unequal variances) was used, while for multiple groups, ANOVA or the Kruskal-Wallis test was used depending on the nature of the distribution. In cases with significant differences ($p < 0.05$), post-hoc analyses were performed using Tukey HSD or Welch comparisons with Bonferroni adjustment, ensuring control of Type I error in multiple

comparisons. Additionally, ordinary least squares (OLS) linear regression models were implemented to perform a Type II ANOVA to evaluate the interaction between variables.

On the other hand, the Wav2Vec 2.0 deep learning architecture [26], developed by Meta AI (Facebook), was integrated to compare automatic detection against human judgment on a balanced sample of 90 audio files (45 authentic and 45 synthetic). The model estimated the complexity of each audio file by calculating Shannon entropy. According to this logic, a clear probability distribution indicated "certainty" in the detection of real audio files (low entropy), while the inconsistencies or uncertainties inherent in a deepfake resulted in greater data complexity (high entropy). The final classification was determined through a calibration phase to find the optimal threshold using the Receiver Operating Characteristic (ROC) curve [27]. The results were inversely normalized on a scale of 0 to 100 to allow for a direct comparison of accuracy between the artificial intelligence and the participants.

3. Results

3.1. Human Performance and Statistical Significance

Regarding the analysis of human performance, it was observed that demographic variables such as gender and nationality did not show statistically significant differences in the evaluated classification metrics. In the case of gender, the performance profiles for men and women are remarkably similar, with minimal variations in precision and specificity that do not establish a robust comparative advantage. Similarly, when comparing Paraguayan citizens to foreigners, the behavior of the accuracy, F1 Score, and TPR metrics remains within close ranges, suggesting that geographic origin is not a determining factor in the discrimination capacity of the analyzed task. Due to this lack of significance, the detailed results for these two categories are omitted from the main graphical representation to prioritize the findings with greater statistical weight.

On the other hand, the analysis revealed that educational level and generational range do exert a significant influence on performance, as shown in Figure 3. In terms of education (Figure 3a), an upward progression is observed, with participants with postgraduate and university studies consistently outperforming those with basic education, especially in accuracy and F1 Score metrics. This indicates that higher academic training translates into more balanced and technical decision-making. Regarding generations (Figure 3b), the results show that Millennials and Generation Z exhibit superior performance compared to Baby Boomers. This gap is particularly visible in the accuracy and specificity metrics, indicating that younger generations are more effective at correctly identifying the evaluated categories, possibly due to greater familiarity with the analyzed digital environments.

Complementarily, the analysis of self-perception versus actual performance reveals a positive and statistically significant relationship between confidence level and accuracy ($p = 0.0238$), although this connection is weak given the Spearman's rank correlation coefficient of 0.1065. The data distribution shows that, while accuracy tends to increase slightly as reported confidence increases, there is a critical dispersion at confidence levels 4 and 5, where cases with significantly low actual accuracy are recorded. This indicates that the low-performing groups do not necessarily operate under a state of uncertainty, but rather exhibit marked overconfidence, reporting maximum levels of certainty despite delivering inaccurate results; this cognitive dissonance invalidates confidence as a quality filter.

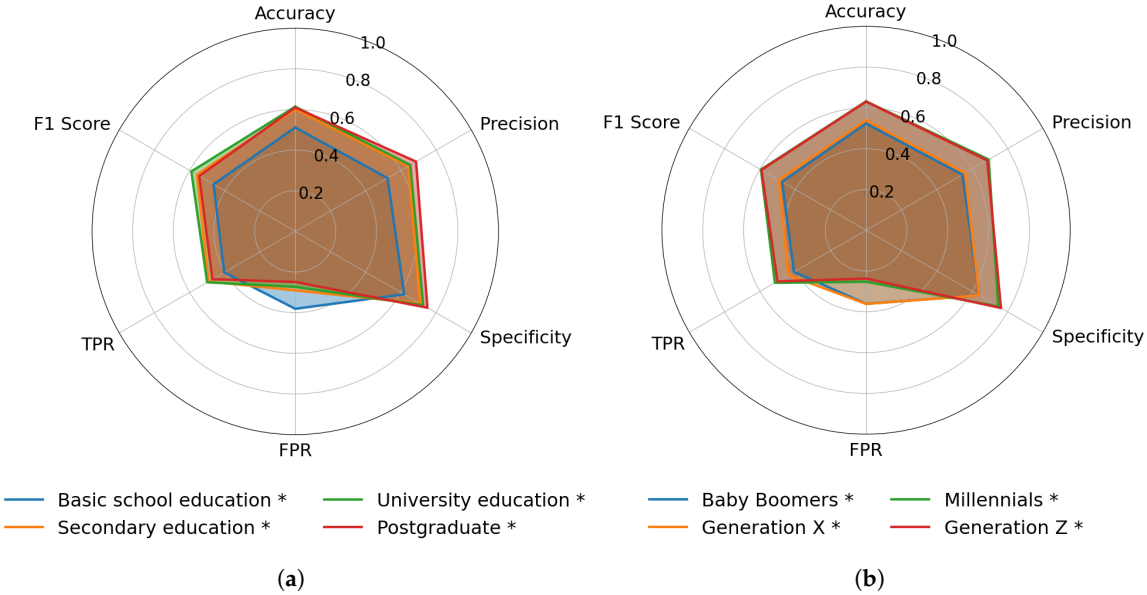


Figure 3. Multidimensional analysis of human performance based on: (**a**) education, (**b**) generation. The asterisks (*) indicate categories that show statistically significant differences ($p < 0.05$).

3.2. Interaction Between the Gender of the Speaker and the Listener

As detailed in Table 3, the results revealed a significant main effect of listener gender ($F = 9.964$, $p < 0.01$), with men showing superior performance. However, no evidence was found of a main effect of speaker gender ($p = 0.598$) or of a significant interaction between listener and speaker gender ($p = 0.411$). This suggests that success in identifying the audio depends on who is listening, but not on the specific gender combination.

Table 3. ANOVA results for gender factors.

Factor	*SS*	*df*	*F*	*p*
Listener Gender (L)	0.38	1.0	9.964	0.002
Speaker Gender (S)	0.01	1.0	0.278	0.598
Interaction (L × S)	0.03	1.0	0.676	0.411

Note. *SS* = Sum of Squares; *df* = degrees of freedom; *F* = F-statistic; *p* = *p*-value. Shading denotes significance.

3.3. Horizontal Comparison of Human Capability vs. AI

Comparative analysis reveals that human evaluators demonstrated a superior and more consistent ability to classify the authenticity of the audio recordings compared to the AI (see Figure 4). While the human group maintained a confidence level close to 60% in both categories—consistently exceeding the level of chance (50%)—the AI model showed significantly lower and more erratic performance. The notable difference in the morphology of the box plots suggests that humans possess a more robust perceptual sensitivity for detecting nuances of naturalness. In the human group, the low dispersion of the data indicates high inter-subject consensus, while the AI exhibits a bimodal or highly skewed distribution, with a large concentration of values in the minimum ranges (0–20%).

The AI data exhibits a marked tendency toward error, especially in the authentic audio category, where its median fell drastically below 20%. This variability indicates that, although the AI architecture can correctly identify certain isolated patterns, it lacks the stochastic stability necessary to discriminate authenticity without generating a high false negative rate. Unlike humans, whose error is more homogeneously distributed around

the median, the AI presents numerous outliers in the synthetic audio category, reaching confidence levels of 80% in isolated cases, but systematically failing in the rest. This suggests that the evaluated model has not managed to encode a generalizable representation of "naturalness", but rather detects non-generalizable technical anomalies.

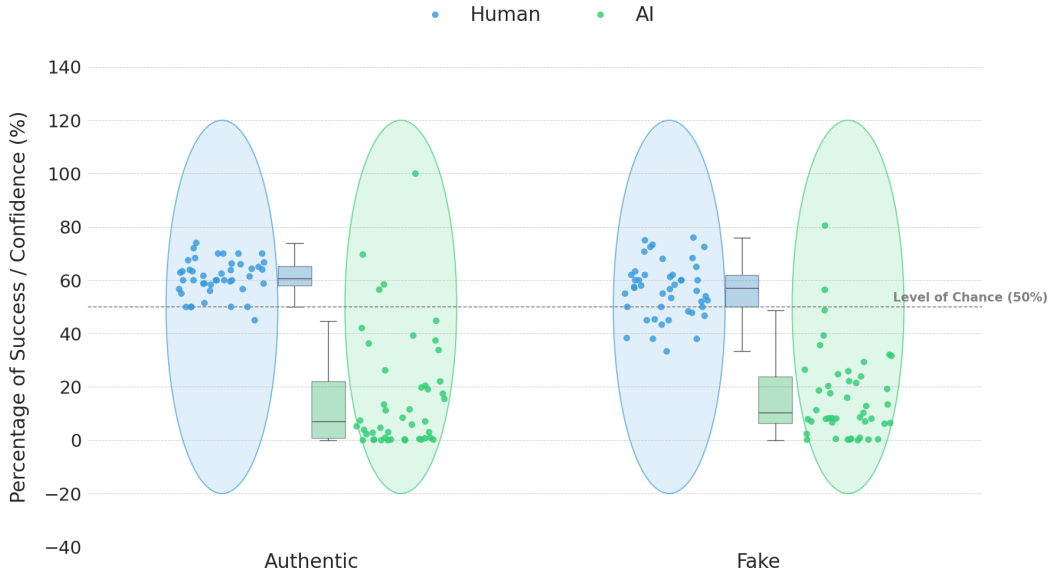


Figure 4. Comparison of perceptual performance between human evaluators and an AI model.

4. Discussion

4.1. Key Findings

Our study demonstrates a significant gap in the perception of vocal authenticity, with humans exhibiting more robust and consistent perceptual sensitivity than artificial intelligence. While the human group consistently performed above the level of chance, the AI showed structural fragility, with a marked tendency toward error and an inability to encode a generalizable representation of "naturalness." This striking difference in the morphology of the results suggests that the human eye—or in this case, ear—detects nuances of fluency that current models still process as isolated technical anomalies.

Analyzing demographic variables, the findings reveal that the ability to discriminate is not universal, but rather mediated by educational level and generational range. The advantage observed in younger generations (Millennials and Generation Z) raises critical questions about the origin of this expertise: is it primarily due to superior digital literacy and prior exposure to synthetic media, or are biological factors related to fine auditory processing, which degrades with age, involved? Similarly, the fact that participants with postgraduate studies outperform those with basic education indicates that academic training equips individuals with analytical tools for more technical decision-making.

A turning point in this analysis is the relationship between performance and confidence indices. It remains to be determined whether lower-performing groups, such as Baby Boomers or people with basic education levels, are simply insecure about the task or, conversely, exhibit overconfidence leading to erroneous judgments. This distinction is vital for understanding how prosodic anomalies—unnatural variations in intonation, rhythm, or micro-segmental pauses—consistently lead to misclassification by humans, especially when the timbre of the synthesized audio is high but it lacks rhythmic expressiveness.

Finally, the lack of statistical significance regarding gender and nationality suggests that the ability to discern "acoustic truth" transcends cultural and identity boundaries, establishing itself as a cross-cutting cognitive skill. However, the ineffectiveness of current automated defenses against the stability of human judgment underscores a technological

urgency: until we succeed in enabling AI to replicate the sophistication of human auditory processing, detection systems will remain insufficient to meet the challenges posed by synthetic speech in today's society. The human ear, for now, is a judge that AI cannot bribe.

4.2. Challenges and Limitations

Based on the findings, this study provides a solid starting point for understanding the factors that influence the detection of counterfeit audio recordings. However, the scope of analysis is strictly limited to audio recordings in a very specific type of Spanish: that spoken in Paraguay. Consequently, the results cannot be generalized to all Spanish speakers, as they focus on a specific group of people, introducing biases that limit the representativeness of the data, limiting the sample size, and requiring caution in extrapolation.

4.3. Future Research Directions

Given the notably limited number of audio deepfake samples available in languages other than English, and given the valuable contribution made in this research with a small dataset specifically generated in Spanish, a priority line of research would be the creation of much more robust and diverse multilingual datasets. In this context, an interesting avenue to explore in future work would be the evaluation of the impact of accents on the successful detection of deepfakes, given that this has not yet been investigated in the scientific literature, constituting an area that requires urgent attention.

A complementary issue to consider would relate to the need to address the validity and fairness of training datasets (biased and unbalanced data) due to the presence of certain attributes that lead to poor performance in terms of detection capacity. Similarly, it could be beneficial to employ a financially incentivized approach to recruiting participants to further explore human behavior by encouraging them to pay full attention and engage more actively and deeply in observational tasks.

However, in addition to the use of more advanced cloning tools, future comparative work would involve conducting two contrasting experiments: one with normal people and the other with sound analysis experts. A detailed study of the waveforms of authentic and fake audio, including a comprehensive multifactorial analysis of audio quality, would also be necessary. Furthermore, one of the most pressing challenges would be the real-time, large-scale detection of fake content, which requires cross-cutting solutions.

Finally, and as a direct extension of the observational analysis, it is considered essential to address the interaction between human variables and the technical characteristics of the audio. Specifically, we propose to investigate whether there are significant interaction effects between the demographic factors of the participants and the properties of the audio that could enhance or mitigate the accuracy of human detection of audio deepfakes. This latter approach will not only enrich our understanding of perception, but will also inform the design of more adaptive awareness and detection systems.

5. Conclusions

The results demonstrate a superior cognitive ability in detecting vocal naturalness that current AI architectures cannot yet replicate, highlighting the lack of stochastic stability in automated models to serve as reliable defenses. The fact that human performance is superior but dependent on educational level and generational familiarity with digital environments underscores the inadequacy of current defenses. These findings emphasize the urgent need to develop new detection tools that encode more robust representations of naturalness and to increase technical training to reduce vulnerability to synthetic speech.

Author Contributions: Conceptualization, M.V.G.R.; methodology, M.V.G.R.; software, M.V.G.R.; validation, M.V.G.R. and P.P.-E.; formal analysis, M.V.G.R. and P.P.-E.; investigation, M.V.G.R.; re-

sources, M.V.G.R., P.P.-E. and E.D.; data curation, M.V.G.R.; writing—original draft preparation, M.V.G.R.; writing—review and editing, M.V.G.R., J.P.-R. and E.D.; visualization, M.V.G.R.; supervision, E.D.; project administration, E.D. All authors have read and agreed to the published version of the manuscript.

Funding: Financial support from the following project is gratefully acknowledged: ANID Doctorado Nacional (/2022-21220867).

Institutional Review Board Statement: The collection of the dataset was approved on 12 February 2025 by the Research Ethics Committee (REC) of the Universidad Sudamericana.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All resources presented in this study are publicly available on Kaggle. The set of 90 MP3 audio files (used in the observational tests by the ReFake application) can be found at: https://doi.org/10.34740/kaggle/ds/5773587. The survey data, consisting of responses from the 450 participants, are available as a CSV file at: https://doi.org/10.34740/kaggle/ds/9480071.

Acknowledgments: The authors wish to express their gratitude to all those who, with their time and expertise, made this research possible. Special thanks to the team of journalists, the 450 participants, and those who enabled the observational group to grow substantially.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Collins, A. *Forged Authenticity: Governing Deepfake Risks*; Technical Report; EPFL International Risk Governance Center (IRGC): Lausanne, Switzerland, 2019. Available online: https://scholar.google.com/scholar_lookup?title=Forged+Authenticity:+Governing+Deepfake+Risks&author=Collins,+A.&publication_year=2019 (accessed on 15 September 2024).
2. Gaur, L. (Ed.) *Deepfakes: Creation, Detection, and Impact*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2023. [CrossRef]
3. Lyon, B.; Tora, M. *Exploring Deepfakes: Deploy Powerful AI Techniques for Face Replacement and More with This Comprehensive Guide*; Packt Publishing: Birmingham, UK, 2023.
4. Gambín, Á.F.; Yazidi, A.; Vasilakos, A.; Haugerud, H.; Djenouri, Y. Deepfakes: Current and future trends. *Artif. Intell. Rev.* **2024**, *57*, 64. [CrossRef]
5. Groh, M.; Sankaranarayanan, A.; Singh, N.; Kim, D.Y.; Lippman, A.; Picard, R. Human detection of political speech deepfakes across transcripts, audio, and video. *Nat. Commun.* **2024**, *15*, 7629. [CrossRef]
6. Bray, S.D.; Johnson, S.D.; Kleinberg, B. Testing human ability to detect 'deepfake' images of human faces. *J. Cybersecur.* **2023**, *9*, tyad011. [CrossRef]
7. Müller, N.M.; Pizzi, K.; Williams, J. Human perception of audio deepfakes. In Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Lisboa, Portugal, 14 October 2022; pp. 85–91. [CrossRef]
8. Stroebel, L.; Llewellyn, M.; Hartley, T.; Ip, T.S.; Ahmed, M. A systematic literature review on the effectiveness of deepfake detection techniques. *J. Cyber Secur. Technol.* **2023**, *7*, 83–113. [CrossRef]
9. Diel, A.; Lalgi, T.; Schröter, I.C.; MacDorman, K.F.; Teufel, M.; Bäuerle, A. Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Comput. Hum. Behav. Rep.* **2024**, *16*, 100538. [CrossRef]
10. Frank, J.; Herbert, F.; Ricker, J.; Schönherr, L.; Eisenhofer, T.; Fischer, A.; Dürmuth, M.; Holz, T. A Representative Study on Human Detection of Artificially Generated Media Across Countries. In *2024 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 55–73. [CrossRef]
11. Yi, J.; Wang, C.; Tao, J.; Zhang, X.; Zhang, C.Y.; Zhao, Y. Audio deepfake detection: A survey. *arXiv* **2023**, arXiv:2308.14970. [CrossRef]
12. Almutairi, Z.M.; Elgibreen, H. Detecting fake audio of arabic speakers using self-supervised deep learning. *IEEE Access* **2023**, *11*, 72134–72147. [CrossRef]
13. Dixit, A.; Kaur, N.; Kingra, S. Review of audio deepfake detection techniques: Issues and prospects. *Expert Syst.* **2023**, *40*, e13322. [CrossRef]
14. Mcuba, M.; Singh, A.; Ikuesan, R.A.; Venter, H. The effect of deep learning methods on deepfake audio detection for digital investigation. *Procedia Comput. Sci.* **2023**, *219*, 211–219. [CrossRef]
15. Xie, Z.; Li, B.; Xu, X.; Liang, Z.; Yu, K.; Wu, M. FakeSound: Deepfake general audio detection. *arXiv* **2024**, arXiv:2406.08052. [CrossRef]

16. Akhtar, Z.; Pendyala, T.L.; Athmakuri, V.S. Video and Audio Deepfake Datasets and Open Issues in Deepfake Technology: Being Ahead of the Curve. *Forensic Sci.* **2024**, *4*, 289–377. [CrossRef]
17. Rabhi, M.; Bakiras, S.; Di Pietro, R. Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Syst. Appl.* **2024**, *250*, 123941. [CrossRef]
18. Alali, A.; Theodorakopoulos, G. Partial fake speech attacks in the real world using deepfake audio. *J. Cybersecur. Priv.* **2025**, *5*, 6. [CrossRef]
19. Ouajdi, H.; Hadder, O.; Tailleur, M.; Lagrange, M.; Heller, L.M. Detection of deepfake environmental audio. In *2024 32nd European Signal Processing Conference (EUSIPCO)*; IEEE: Lyon, France, 2024; pp. 196–200. [CrossRef]
20. Mai, K.T.; Bray, S.; Davies, T.; Griffin, L.D. Warning: Humans cannot reliably detect speech deepfakes. *PLoS ONE* **2023**, *18*, e0285333. [CrossRef] [PubMed]
21. Rathgeb, C.; Tolosana, R.; Vera-Rodriguez, R.; Busch, C. (Eds.) *Handbook of Digital Face Manipulation and Detection From DeepFakes to Morphing Attacks*; Springer: Berlin/Heidelberg, Germany, 2022. [CrossRef]
22. Babaei, R.; Cheng, S.; Duan, R.; Zhao, S. Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis. *J. Sens. Actuator Netw.* **2025**, *14*, 17. [CrossRef]
23. Han, C.; Mitra, P.; Billah, S.M. Uncovering human traits in determining real and spoofed audio: Insights from blind and sighted individuals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2024; pp. 1–14. [CrossRef]
24. Cooke, D.; Edwards, A.; Barkoff, S.; Kelly, K. As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli. *arXiv* **2024**, arXiv:2403.16760. [CrossRef]
25. Tharwat, A. Classification Assessment Methods. *Appl. Comput. Inform.* **2021**, *17*, 168–192. [CrossRef]
26. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460. Available online: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html (accessed on 12 January 2026).
27. Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.