FORUM ACUSTICUM EURONOISE 2025

PERCEPTION OF AUDIO DEEPFAKES IN SPANISH AND JAPANESE: EFFECTS OF LANGUAGE, SPEAKING STYLE, AND VOICE FAMILIARITY

Aurora López-Jareño[1*] Juan José Romero[1] Eugenia San Segundo[1] Xin Wang[2]
[1] Phonetics Laboratory, Spanish National Research Council (CSIC), Madrid, Spain
[2] National Institute of Informatics, Tokyo, Japan

ABSTRACT

Deepfakes are posing significant challenges to forensic phonetics, undermining citizen security and trust in digital media. Thus, understanding the human ability to distinguish synthetic audio from authentic audio is crucial in addressing this growing threat.
Using PsychoPy, we conducted a perceptual experiment in which participants classified real and fake audio samples. The test featured Spanish and Japanese stimuli distributed to Spanish native speakers to examine the impact of language knowledge on performance. [1-2] have explored this variable, whose results we aim to compare with our findings. Additionally, this study evaluates how speaking style (interviews vs. text reading) and familiarity with the speaker's voice impact performance.
The experiment includes 80 voice samples (M=10.15 s), 50% real and 50% fake. For the real interview samples, we selected 10 Spanish stimuli from VoxCeleb-ESP [3] and 10 Japanese stimuli from EACELEB [4]. For the 20 real text-reading samples, 20 Spanish and Japanese were sourced from LibriVox and YouTube audiobooks. Furthermore, these 40 real stimuli (interviews and text reading) were cloned using Eleven Labs to generate their synthetic counterparts.

Keywords: *deepfakes, forensic phonetics, voice perception, Spanish, Japanese.*

[*] ***Corresponding author***: *aurora.lopez@cchs.csic.es*

1. INTRODUCTION

Audio deepfakes are AI-generated or AI-edited speech that closely resembles real human speech [5]. While this technology offers potential benefits in clinical applications, like voice cloning for patients with neurodegenerative diseases, it also poses significant risks to society. Deepfakes can make individuals more vulnerable to misinformation and manipulation, as synthetic audio or video content can be difficult to distinguish from authentic material. Furthermore, not only humans but also security systems based on voice recognition are susceptible to deepfake attacks, increasing the risk of spoofing attempts [6].
In this context, improving deepfake detection technologies and raising public awareness about their existence and potential threats is essential. Understanding human perception of deepfake voices is particularly relevant, as it helps assess how vulnerable people are to synthetic speech. Moreover, perceptual studies can provide insights into the linguistic and phonetic differences between real and deepfake voices, which could, in turn, enhance automatic detection systems by incorporating new discriminative parameters. Despite its importance, human detection capability has been little explored, and most perceptual studies have focused on the English language.
Our study aims to contribute to the understanding of human ability to distinguish real human voices from deepfake voices in Spanish and Japanese, considering various factors that may influence detection accuracy. Specifically, we address the following research questions:

(1) Does language familiarity influence the ability to distinguish deepfake voices from real human voices?
(2) Does speaking style affect the accuracy of voice discrimination?

(3) Does prior familiarity with a speaker's voice improve the ability to detect cloned voices?

Thus, the main variables studied in this work are language proficiency, speaking style and familiarity with the speaker's voice.

Language proficiency has been studied in previous research with varying conclusions. On the one hand, [1] found that native English speakers recognized English deepfake audio slightly better than non-native speakers, a result also observed by [7]. On the other hand, [2] conducted an experiment with English and Mandarin native speakers and did not observe differences in human detection accuracy based on language proficiency.

Regarding speaking style, previous perceptual studies, such as [2, 8], have primarily focused on text-reading recordings, rather than spontaneous speech. However, we considered it relevant to compare speaking styles with different degrees of spontaneity in order to fill this research gap. Spontaneous speech contains a wide range of phenomena that are challenging for AI models to replicate, such as hesitations, filled pauses, clicks, truncated or incomplete words, among others [9]. Thus, artificial interview stimuli will likely seem more unnatural due to difficulty of cloning some features associated with spontaneous speech.

Regarding voice familiarity, neuroscientific studies have observed differences in neural processing between familiar and unfamiliar natural voices. For instance, [10] found that greater familiarity is associated with larger BOLD signal amplitudes in the temporal lobes. However, perceptual studies on deepfake detection that consider this variable remain scarce. In a previous study, a survey of 200 participants indicated that humans can distinguish artificial from real voices with approximately 50% accuracy when the voices are unfamiliar, but this rate increases to around 80% when the voices are familiar [11]. The present study aims to shed light on this largely unexplored relationship between voice familiarity and deepfake detection, addressing a notable gap in the current literature.

In summary, based on our literature review, we formulate the following hypotheses:

(1) *Language familiarity improves detection accuracy.*
(2) *Deepfake interview stimuli are easier to identify than audiobooks stimuli.*
(3) *A deepfake of a familiar voice is easier to identify than an anonymous voice.*

The remainder of this article is structured as follows: Section 2 details the methodology, Section 3 outlines the results, and Section 4 discusses them and presents our conclusions.

2. METHODOLOGY

2.1. Survey Design

Using the open-source software PsychoPy [12], we designed a phonetic perceptual experiment lasting approximately 25 minutes, which was then uploaded to Pavlovia for online distribution. The experiment could be completed in various electronic devices (PC, tablets, smartphone) without any time limitation. It aimed to assess human ability to distinguish between real human voices and their deepfake counterparts. Thus, participants were exposed to 80 audio clip stimuli presented in a randomized order, with an equal distribution of natural and artificial clips (one synthetic stimuli per counterpart: 40 natural and 40 deepfake). After listening to each clip, participants were required to classify it as either "natural", if they believed it was a real human voice, or "artificial", if they thought it was AI-generated.

Additionally, after each response, they rated their confidence level on a 5-point Likert scale, where 1 indicated 'not at all confident' and 5 indicated 'completely sure'. Participants did not receive any feedback on their performance during the experiment, nor were they informed about the proportion of artificial to natural stimuli.

The experiment was divided into two parts, with a break in between. In the first part, the clips were extracted from audiobooks read in Spanish and Japanese. In the second part, the clips featured Spanish and Japanese celebrities speaking in interviews. This second part included two additional tasks: a yes/no question about whether they knew the celebrity's voice and an open-ended question asking them to explain the reasoning behind their 'natural'/'artificial' classification.

The distribution of stimulus types is identical in both parts of the experiments, as described in Table 1.

Table 1. Distribution of audio stimuli by nature, language, and speaker sex in both experimental parts.

Nature	*Language*	*Speaker sex*
20 natural	10 Japanese	5 female
		5 male
	10 Spanish	5 female
		5 male
20 artificial	10 Japanese	5 female
		5 male
	10 Spanish	5 female
		5 male

2.2. Participants

The final dataset comprised 2,211 responses from 28 native Spanish listeners (50% male and 50% female), recruited via the Pavlovia server. The participants' age ranged from 22 to 65 years (M_{age}=30.9 years, SD = 10.34). Additional potentially relevant characteristics were collected, including Japanese language proficiency (rated on a Likert scale from 'null' to 'high') and advanced linguistic knowledge (self-reported academic background)[1]. The distribution of these characteristics was as follows: regarding linguistic expertise, 39.29% of participants reported having specific training; regarding Japanese proficiency, 24 participants reported no competence, 3 reported low competence, and 1 reported an intermediate level. Non-native Spanish speakers and individuals with hearing impairments were excluded from the study. All participants provided informed consent and were said to complete the experiment in a quiet environment using headphones.

2.3. Stimuli Selection

We selected 80 audio stimuli with an average duration of 10.15 seconds (Mdn = 10.04 s), ranging from 6 to 12 seconds. For each natural stimulus, we synthesized its synthetic counterpart, maintaining the exact same phrase as in the natural voice clip. All stimuli were processed into MP3 format, with a 44,100 Hz sampling frequency and in a single-channel output.

2.3.1. Bonafide stimuli

To obtain 20 real text-reading samples, we sourced ten Spanish and ten Japanese audiobooks from LibriVox and YouTube. Each stimulus was extracted from a different audiobook read by a different speaker. Then, the software Praat [13] was used to trim a 10-second fragment from each full recording.

The 10 stimuli of Spanish celebrity interviews were obtained from the VoxCeleb-ESP corpus [3]. The selected Spanish celebrities represented a broad spectrum of public figures, including singers, journalists, television hosts, actors, athletes and comedians. Additionally, we aimed to include celebrities from various Spanish regions to capture geographic accent diversity. On the other hand, the 10 stimuli from Japanese celebrity interviews were sourced from the EACELEB corpus [4]. In this case, all selected celebrities were either actors or singers, and regional accent diversity could not be ensured. For both corpora, we established the following exclusion criteria before selecting the stimuli:

(a) Presence of background noise or music
(b) Poor recording quality
(c) Interruptions by the interviewer or audience
(d) Insufficient material in the corpus to generate a cloned voice
(e) Presence of political and controversial content

The last criterion was included to minimize extraneous cues in the content, as the experiment aimed to evaluate phonetic characteristics that can be used to distinguish real voices from fake voices.

Gender balance was maintained, resulting in an equal distribution of male and female voices for both Spanish and Japanese audiobook and interview stimuli (see Table 1).

2.3.2. Voice Cloning

We ensured that each synthetic voice reproduced the exact same phrase as its natural counterpart. This allowed for a direct comparison between real and artificially generated voices, isolating perceptual differences to the phonetical characteristics themselves, rather than linguistic content. Thus, transcriptions of the natural stimuli were necessary to generate deepfakes that precisely replicated the original audio clips. The automated transcription tool Whisper was utilized for this purpose [16]. Subsequently, the transcriptions underwent a rigorous review and correction process by two qualified linguists.

To produce synthetic versions of the natural voices, we used ElevenLabs' Text-to-Speech (TTS) software [14]. We applied the "Eleven Multilingual v2" model with its default settings for Stability, Similarity, Style Exaggeration, and Speaker Boost, adhering to ElevenLabs' "Best Practices" guidelines [15] and the ElevenLabs Prompt Guide for inputting both audio and text.

[1]The criterion for considering a participant to have advanced linguistic knowledge was that they were at least enrolled in an undergraduate degree in Linguistics.

For voice cloning, each target voice required a training audio sample of 1 to 2 minutes in duration. These training audio samples were extracted from the same corpora detailed in section 2.3.1, ensuring that the selected experimental stimuli were excluded.
After training the ElevenLabs software and obtaining accurate transcriptions for each natural audio clip, the generation of artificial voices was conducted through an iterative process. For each voice, a minimum of three cloned versions were produced. In certain instances, adjustments to the transcriptions were made during this process to enhance the naturalness of the generated voices, necessitating multiple iterations. Finally, a single deepfake was selected for each voice through a consensus-based approach involving at least three researchers or native-speaking collaborators.

2.4. Data Analysis

For the descriptive analysis, the accuracy rate (%) and standard deviation were calculated per participant using the Equation (1). True positive (TP) refer to the number of artificial stimuli correctly identified as artificial; true negative (TN) represent the number of natural voices correctly classified as natural; false positive (FP) occur when natural voices are misclassified as artificial, and false negatives (FN) refer to artificial voices incorrectly classified as natural.

$$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \quad (1)$$

Equation (1) was also used to compute the mean correct answer rate (%) per audio stimulus, which was then grouped according to different variables of interest.
For the inferential analysis, RStudio [17] was used, employing the lme4 package to construct generalized linear mixed models using the glmer function [18], and the ggplot2 package for data visualization [19]. Two separate regression models were designed:

- Model 1 included the entire dataset.
- Model 2 focused specifically on Spanish interview stimuli, where the variable "voice familiarity" was analyzed. A separate model was necessary for this condition, as familiarity with a speaker's voice was only assessed for Spanish interviews.

In both models, the dependent variable was the probability of a correct response (binary outcome: correct vs. incorrect). The fixed effects included:

- Authenticity (natural vs. artificial),
- Language (Spanish vs. Japanese),
- Confidence in response (mean rating on a 5-point scale),
- Speaker's sex (male vs. female),
- Participant's gender,
- Participant's linguistic background.

Additionally, Model 1 incorporated speaking style (audiobooks vs. interviews), while Model 2 included voice familiarity.
The effect of age was not included due to the uneven distribution of participants across age groups. Although the sample covered a broad age range (22–65 years), the median age (IQR) was 27 (7), indicating a skew toward younger participants. This imbalance likely limited the statistical power to detect age-related effects on the ability to discriminate between natural and deepfake voices.
For random effects, both models included intercepts for participants and audio clips, accounting for individual variability in performance and stimulus difficulty.
To assess the explanatory power of the generalized linear mixed-effects model, we computed the coefficient of determination (R^2), using the r2_nakagawa function from the *performance* R package. This method provides two R^2 values: the marginal R^2, which reflects the variance explained by the fixed effects alone, and the conditional R^2, which accounts for the variance explained by both fixed and random effects.
Furthermore, visual inspection of residual plots confirmed there were no obvious deviations from homoscedasticity.

3. RESULTS

3.1 Overall Participants' Performance

The average accuracy rate of the participants for the total number of audios listened to was 60.2% (SD=8.2%). Participants performed better on Spanish stimuli (M=65.8%) compared to Japanese stimuli (M=55.2%), which was close to chance level performance.
As shown in Table 2, participants achieved higher accuracy when identifying interview stimuli compared to audiobooks.

Table 2. Mean accuracy rate and SD (%) of participants in audio clips discrimination tasks, grouped by speaking style, language, and authenticity. Note. 'A'='artificial'; 'N'='natural'.

Audiobooks				**Interviews**			
55.1 (8.3)				66.0 (10.6)			
Japanese		**Spanish**		**Japanese**		**Spanish**	
52.7 (11.1)		57.5 (10.6)		57.7 (15.1)		74.5 (12.2)	
N	**A**	**N**	**A**	**N**	**A**	**N**	**A**
51.8 (17.7)	53.6 (18.3)	62.6 (12.8)	51.4 (21.6)	65.5 (20.5)	48.9 (22.3)	77.5 (17.4)	71.1 (15.7)

Furthermore, in all cases, participants performed worse in identifying Japanese stimuli than Spanish stimuli. This difference was particularly pronounced in interviews (74.5% for Spanish vs. 57.7% for Japanese). In fact, accuracy rates for Japanese interviews were more similar to audiobooks than to Spanish interviews.

Additionally, participants performed better at identifying natural voices than artificial ones, except for Japanese audiobooks (where artificial voices had slightly higher accuracy: 53.6% vs. 51.8%). On the other hand, the high standard deviation (above 20 in some cases) indicates significant variation in intersubject performance.

3.2 Hypothesis testing

Table 3 summarizes the model 1 fitting results of all predictors. The reference category for the response variable is artificial Spanish audiobooks read by female speakers and answered by male participants with no linguistic knowledge and a confidence level of 1. The model explained 8% of the variance through fixed effects (marginal $R^2 = 0.080$), and 18,4% when including both fixed and random effects (conditional $R^2 = 0.184$), indicating a substantial contribution of subject and item-level random variability.

The results of model 1 confirmed our first two hypotheses. First, Japanese stimuli significantly decreased the probability of accurate classification ($\beta = -0.38$; $p<0.05$), supporting the hypothesis that language familiarity enhances participants' ability to distinguish deepfake voices from real human voices. Second, regression model 1 indicated that interview speaking style significantly increased the probability of correct identification ($\beta=0.49$; $p<0.05$), suggesting that deepfake interviews are easier to identify than deepfake audiobooks.

Additionally, two more variables were significant. Natural stimuli significantly increased the probability of correct identification ($\beta=0.37$; $p<0.05$). Higher participant confidence levels were also associated with better identification accuracy, specifically at confidence level 4 ($\beta = 0.94$; $p < 0.01$) and level 5 ($\beta = 1.48$; $p < 0.001$).

Table 3. Summary of estimated regression parameters for model 1: Estimate, standard error (SE), z-ratio and p-value. Note. Significance level: $p < 0.001$***; $p < 0.01$**, $p < 0.05$*

	Estimate	*SE*	*z value*	*p*
Intercept	−0.51	0.39	−1.285	0.20
Authenticity: natural	0.37	0.17	2.247	0.02 *
Language: Japanese	−0.38	0.17	−2.237	0.03 *
Speaking style: interview	0.49	0.17	2.955	0.003 **
Speaker sex: male	0.01	0.17	0.036	0.97
Confidence level: 2	0.52	0.36	1.461	0.14
Confidence level: 3	0.53	0.35	1.521	0.13
Confidence level: 4	0.94	0.35	2.676	0.007 **
Confidence level: 5	1.48	0.38	3.919	< 0.001 ***
Linguistics knowledge: yes	0.25	0.13	1.910	0.06
Gender listener: female	−0.22	0.13	−1.792	0.07

The box plots in Figure 1 help visualize how the predictor variables identified in the regression model 1 influence accuracy performance. As it is shown, natural stimuli have higher correct answer rates than their counterparts (e.g. Spanish natural interviews vs. Spanish artificial interviews vs.). This means that natural stimuli are easier to identify than artificial voices, so there are more false negatives than false positives.

As box plots show, regarding language, the medians of Spanish voices are higher than Japanese voices in almost all cases (Fig. 1). Only Japanese artificial audiobooks showed a slightly higher correct answer median compared to Spanish ones.

Concerning speaking style, Spanish artificial interviews were identified with higher accuracy (median accuracy close to 75%) compared to Spanish artificial audiobooks (below 50%), suggesting a higher rate of false negatives in the latter condition. However, Japanese artificial interviews

and Japanese artificial audiobooks are similar in correct answer rate.

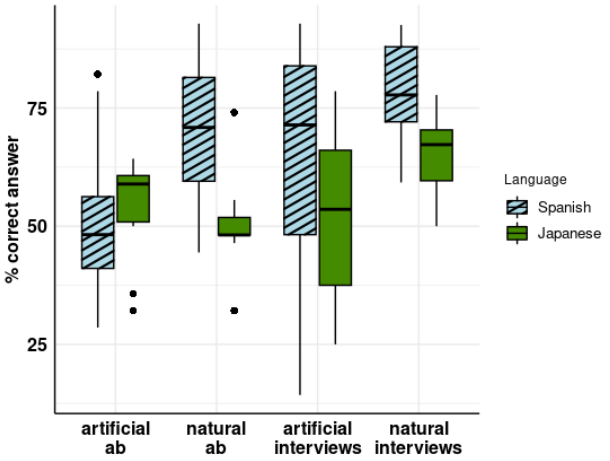

Figure 1. Box plots of mean correct answer rate (%) per audio stimulus, grouped by language (Spanish vs. Japanese), speaking style (audiobooks vs. interviews) and authenticity (natural vs. artificial). Note. 'Ab' = audiobooks.

On the other hand, a separate regression model focused on Spanish interview stimuli (model 2) examined the effect of voice familiarity on identification accuracy. This model explained 15,1% of the variance through fixed effects (marginal R^2 = 0.151), and 19,7% when including both fixed and random effects (conditional R^2 = 0.197).
The results of model 2 indicate that voice familiarity did not significantly influence deepfake recognition. This finding contradicts our third hypothesis, as data suggested that knowing the original voice of a speaker does not substantially improve deepfake detection accuracy. The mean accuracy for familiar and unfamiliar voices were similar (76.7% vs. 73.9%), although standard deviation for unfamiliar voices was notably higher (30.8 %), indicating greater variability across participants.

3.3. Confidence

As confirmed by the regression models, confidence was a significant predictor of response accuracy, emerging as the most statistically relevant variable. The scatter plots in Figure 2 further illustrate how confidence influences the likelihood of correct responses.
For the Japanese stimuli, no clear correlation was observed between confidence and accuracy (Fig. 2). Confidence ratings were generally lower compared to Spanish stimuli, which aligns with the fact that most participants were native Spanish speakers with no knowledge of Japanese. Notably, only one Japanese audio sample had an average confidence rating exceeding 3.5. Accuracy distribution showed less dispersion than in the Spanish stimuli.
In contrast, for the Spanish stimuli, a positive correlation emerges for confidence levels above approximately 3.75, where higher confidence ratings were associated with higher accuracy rates. Below a confidence level of 3.6, accuracy varied widely across audio samples: in some cases, with some below 50%, others around 50%, and some exceeding 75%. Additionally, the Spanish scatterplot exhibited greater dispersion in accuracy levels, suggesting variability in how confidence relates to performance.
These findings suggest that language familiarity plays a key role in confidence judgments and their relationship with accuracy.

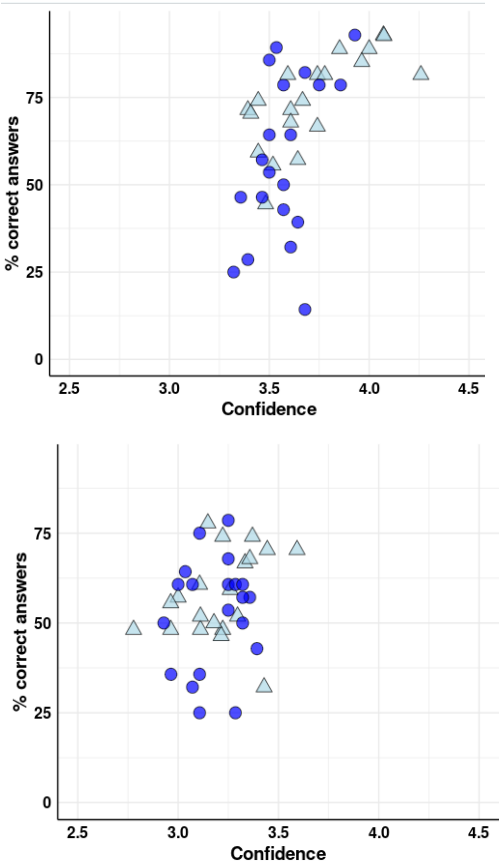

Figure 2. Scatterplots showing the relationship between average confidence and mean correct answer rate (%) for each audio stimulus. Top: Spanish; bottom: Japanese. Triangles: interviews; circles: audiobooks.

4. DISCUSSION & CONCLUSION

Multiple implications can be drawn from the results of this study.

First, overall participants' performance (60.2%) lies between previously reported results: lower than [1-2] (approx. 70%), but higher than [7] (53.7%), whose experiment was conducted under more realistic, "in the wild", conditions. This supports the idea that detection accuracy decreases in less controlled environments. Consistently, recent studies report that participants are deceived by audio deepfakes up to 87% of the time in real-world contexts [20].

Secondly, as in [7], participants in our study were more accurate at identifying authentic than synthetic voices, suggesting a bias toward classifying clips as real. Given that they were explicitly informed about the presence of deepfakes—a condition not typically present outside the lab—the actual false negative rate in real-world scenarios may be even higher, highlighting the societal risks posed by AI-generated voices.

Regarding the low detection accuracy in the Spanish audiobook condition, a likely explanation is the high quality of voice cloning in these samples. This may be linked to the *speaking style* variable: audiobooks involve scripted, less spontaneous speech, with fewer prosodic irregularities that are typically harder to replicate. Consequently, TTS algorithms may yield more convincing results when cloning audiobooks than interviews. This interpretation aligns with previous research identifying prosodic modelling as a key challenge for TTS systems [21]. These difficulties can make synthetic voices more detectable—particularly in spontaneous speech contexts—while more structured speech, like that in audiobooks, may reduce these detectable mismatches.

Another factor worth considering is the potential variation in the level of development of the ElevenLabs TTS system across different languages. Thus, the differences observed between Spanish and Japanese conditions may not solely reflect listeners' perceptual abilities, but also uncontrolled variables such as the quality of synthetic voice generation in each language. Japanese audiobooks may have been less accurately cloned by ElevenLabs, making the synthetic speech easier to detect. This could explain why participants achieved higher accuracy rates for Japanese artificial audiobooks compared to their Spanish counterparts.

In third place, this study indicated that there was no significant correlation between accuracy and participants' gender or linguistic background. In future studies it would be interesting to collect a larger and more homogeneous sample to analysis additional demographic factors such as participant age or listener's musical training [22].

Moreover, unlike the findings reported in [11], the present study did not find a significant effect of voice familiarity on participants' ability to detect deepfakes. Given the novelty of this research area, further investigation is required. Future studies could explore different degrees of familiarity, including synthetic versions of participants' own voices, as well as voices of acquaintances, friends, and public figures. Studying celebrity deepfakes could provide insights into our vulnerability to fake news, election manipulation, or defamation campaigns targeting public figures. On the other hand, investigating the perception of deepfakes involving acquaintances may be particularly relevant for understanding susceptibility to scams such as vishing.

Other avenues for future research emerge from the present study. (a) It would be valuable to evaluate the in-domain and out-of-domain performance of an algorithm trained on the voices used in this experiment, and to compare its detection accuracy with that of human participants, following a similar approach to that of [2]. (b) Further exploration of reaction times in deepfake discrimination, as proposed by [23], could clarify whether speed correlates with accuracy. (c) A qualitative analysis of participants' open-ended responses may shed light on the perceptual cues humans rely on to differentiate cloned from authentic voices, in line with [2]. (d) Beyond perceptual cues, examining the acoustic features of natural and synthetic stimuli—as in [24]—could reveal measurable indicators linked to classification accuracy.

5. ACKNOWLEDGMENTS

Grant PID2021-124995OA-l00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

6. REFERENCES

[1] N. Müller, K. Pizzi, and J. Williams, "Human Perception of Audio Deepfake," in *Proc. of the 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM '22),* (Lisboa, Portugal), pp. 85-91, 2022.

[2] K. T. Mai, S. Bray, T. Davies, and L. D. Griffin: "Warning: Humans cannot reliably detect speech deepfakes," *PLoS One*, vol. 18, no. 8, pp. 111–222, 2023.

[3] B. Labrador, M. Otero-González, A. Lozano-Diez, D. Ramos, D. T. Toledano, and J. González-Rofríguez:

"Voxceleb-ESP: preliminary experiments detecting Spanish celebrities from their voices," *Preprint from arxiv*, 2023.

[4] D. Caulley, Y. Yang, and D. Anderson: "Eaceleb: An east asian language speaking celebrity dataset for speaker recognition," *Preprint from arxiv*, 2022.

[5] Z. Khanjani, G. Watson, and V. P. Janeja: "Audio deepfakes: A survey," *Frontiers in Big Data*, vol. 5, pp.1-24, 2023.

[6] E. San Segundo: *La fonética forense. Nuevos retos y nuevas líneas de investigación*. Barcelona: Ediciones Octaedro, 2023.

[7] Di Cooke, A. Edwards, S. Barkoff, and K. Kelly: "As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli," *Preprint from arxiv*, 2025.

[8] G. Watson, Z. Khanjani, and V. P. Janeja: "Audio deepfake perceptions in college going populations," *Preprint from arxiv*, 2022

[9] K. McDougall and M. Duckworth: "Profiling fluency: An analysis of individual variation in disfluencies in adult males," *Speech Communication*, vol. 95, pp. 16-27, 2017.

[10] A. Bethmann, H. Scheich, and A. Brechmann: "The temporal lobes differentiate between the voices of famous and unknown people: an event-related fMRI study on speaker recognition," *PloS one*, vol. 7, no. 10, 2012.

[11] E. Wenger, M. Bronckers, C. Cianfarani, J. Vryan, A. Sha, H. Zheng, and B. Y. Zhao: "'Hello, It's Me': Deep Learning-based Speech Synthesis Attacks in the RealWorld," in *Proc. of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, (virtual event, Republic of Korea), pp. 235-251, 2021.

[12] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv: "PsychoPy2: Experiments in behavior made easy," Behav Res Methods, vol. 51, no. 1, pp. 195–203, 2019

[13] Boersma, Paul & Weenink, David. *Praat: doing phonetics by computer* [Computer program]. Version 6.4.27. http://www.praat.org/

[14] ElevenLabs: Free Text to Speech & AI Voice Generator. https://elevenlabs.io/.

[15] ElevenLabsPrompting-ElevenLabs Documentation. https://elevenlabs.io/docs/best-practices/prompting/controls.

[16] Whisper Transcribe. https://www.whispertranscribe.com/

[17] Posit team: *RStudio: Integrated Development Environment for R* [computer program]. Version 2024.12.1+563. Boston: Posit PBC, 2024. http://www.posit.co/

[18] D. Bates, M. Mächler, B. Bolker, and S. Walker: "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol.67, no.1, pp. 1-48, 2015.

[19] H. Wickham: *ggplot2. Elegant Graphics for Data Analysis.* New York: Springer-Verlag, 2016.

[20] M. Groh, A. Sankaranarayanan, N. Singh, D. Y. Kim, A. Lippman, and R. Picard: "Human detection of political speech deepfakes across transcripts, audio, and video," *Nature Communications*, vol. 15, no. 7629, pp. 1-16, 2024.

[21] J. Yi: "Audio Deepfake Detection: A Survey," *Preprint from arxiv*, 2023.

[22] E. San Segundo: "El entrenamiento musical y otros factores que pueden influir en el reconocimiento perceptivo de hablantes," *Fonética experimental, educación superior e investigación*, vol. 1, pp. 571–588, 2014.

[23] E. San Segundo and M. Gibson, "Reconocimiento perceptivo de hablantes: un experimento con voces clonadas artificialmente y con voces de gemelos idénticos," in *Proc. of the 13th Congreso Ibérico de Acústica Tecniacústica (DDAM '22),* (Faro, Portugal), pp. 1-10, 2024.

[24] E. San Segundo and J. Delgado, "Deepfakes frente a voces de gemelos idénticos: un análisis acústico preliminar," in *Proc. of the 13th Congreso Ibérico de Acústica Tecniacústica (DDAM '22),* (Faro, Portugal), pp. 1-12, 2024.