Comparing humans and neural networks in speech deepfake detection

Manuel Otero-Gonzalez[1,2], *Aurora López-Jareño*[1], *Eugenia San Segundo*[1] *and Joaquin Gonzalez-Rodriguez*[2]

[1]*Phonetics Laboratory*, *Spanish National Research Council (CSIC), Madrid, Spain*

{manuel.otero|aurora.lopez|eugenia.sansegundo}@cchs.csic.es

[2]*AUDIAS, Universidad Autónoma de Madrid, Spain*

joaquin.gonzalez@uam.es

Audio deepfakes generated by modern text-to-speech systems have reached a level of realism that challenges both human listeners, whose accuracy in recent perceptual studies approaches chance levels (Lavan et al., 2025), and automatic detection systems, which often generalize poorly out of domain. This poses significant challenges for forensic phonetics, where speech is used as evidence in forensic speaker comparison (FSC) and authentication tasks.

The few studies comparing automatic detectors and human consensus judgments report comparable performance under realistic conditions (Mai et al., 2023; Warren et al., 2024), with collective human performance matching out-of-domain models, although humans and machines rely on different cues and misclassify different audio samples (Warren et al., 2024).

In this work, we evaluate ResNet18, a state-of-the-art residual deep neural network, using Spanish and Japanese audio samples to compare its performance with human listeners. We hypothesized that a hybrid human-machine approach could enhance classification accuracy. To test this, human and model scores were converted to log-likelihood ratios (LLRs) and both were combined into a hybrid human-machine system using logistic regression. The three systems were further evaluated by measuring their calibration through the log-likelihood ratio cost function (C_{llr}). Finally, to refine the comparative analysis, we qualitatively analysed the participants' reasons for their classification decisions to identify salient characteristics perceived in the audio samples based on concordant and discordant classifications between humans and ResNet18.

Methodology

To evaluate human and model performance, a multispeaker corpus was developed (San Segundo et al., 2025), consisting of 40 natural samples from audiobooks and media interviews in Spanish and Japanese, and 40 synthetic counterparts cloned using the ElevenLabs TTS Multilingual v2 model.

In total, 54 individuals (28 native Spanish and 26 Japanese speakers) were recruited for the detection task. For the automatic detection of deepfakes, we used a ResNet18 (residual network) architecture developed by one of the participating systems (Rohdin et al., 2024) in the ASVspoof5 challenge (Wang et al., 2024). This system was trained and validated on the ASVspoof 5 dataset and subsequently evaluated on the dataset employed in this study.

For metrics comparison, the human performance was measured collectively. The receiver operating characteristic (ROC) curves and their corresponding area under the curve (AUC) were calculated for humans and ResNet18 using the same 80 samples. Additionally, systems were calibrated using logistic regression employing a cross-validation approach (Ramos et al., 2011), and the fusion approach combined both LLRs using the same method

For the qualitative analysis, the methodology of San Segundo et al. (2025) was applied, following the speech classification proposed by Leeman et al. (2024), which is commonly used in FSC.

Results

The hybrid system human-machine showed better performance and the best calibration compared to the individual systems (Figure 1A), with an AUC of 0.8413, C_{llr} of 0.7078 and min-C_{llr} of 0.6867.

This suggests that despite the superior performance of the hybrid system, deepfake detection remains challenging for humans, machines, and fused systems alike.

Following the qualitative study (Figure 1B), we identified perceptual features that stand out particularly in some of the following situations: in audio clips correctly classified by both human listeners and ResNet18 (pattern A), by humans only (pattern B), and by ResNet18 only (pattern C). The preliminary results suggest that laughter, emotion, segmental features, pitch anomalies, and a monotonous intonation/rhythm/intensity/speed/tone may serve as key discriminative clues.

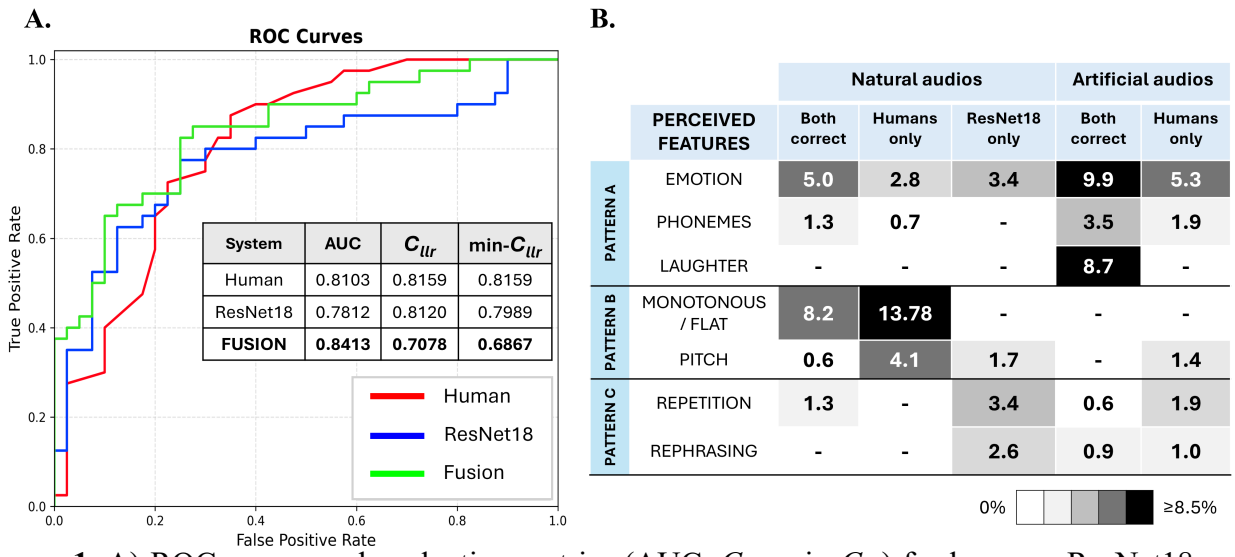

Figure 1. A) ROC curves and evaluation metrics (AUC, C_{llr}, min-C_{llr}) for humans, ResNet18, and the hybrid system (N = 80 audio samples); B) Comparative analysis of human-perceived features across classification outcome groups. Note: Subsets are categorized by whether the human and the model reached the same correct conclusion or diverged. Shading intensity corresponds to normalized frequency levels as indicated in the legend.

References

Lavan, N., Irvine, M., Rosi, V., & Mc Gettigan, C. (2025). Voice clones sound realistic but not (yet) hyperrealistic. *PLoSOne*, *20*(9), e0332692. https://doi.org/10.1371/journal.pone.0332692

Leemann, A., Perkins, R., Buker, G. S., & Foulkes, P. (2024). *An Introduction to Forensic Phonetics and Forensic Linguistics.* Routledge.

Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. PLOSONE, 18(8), e0289567. https://doi.org/10.1371/journal.pone.0285333

Ramos, D., Franco-Pedroso, J., & Gonzalez-Rodriguez, J. (2011). Calibration and weight of the evidence by human listeners: The ATVS-UAM submission to NIST HUMAN-aided speaker recognition 2010. Proceedings of the IEEE ICASSP, 5908–5911. https://doi.org/10.1109/ICASSP.2011.5947706

Rohdin, J., Zhang, L., Oldřich, P., Staněk, V., Mihola, D., Peng, J., Stafylakis, T., Beveraki, D., Silnova, A., Brukner, J., & Burget, L. (2024). BUT systems and analyses for the ASVspoof 5 Challenge. Proceedings of the Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 24–31. https://doi.org/10.21437/ASVspoof.2024-4

San Segundo, E., López-Jareño, A., Wang, X., & Yamagishi, J. (2025). Human perception of audio deepfakes: The role of language and speaking style [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2512.09221

Wang, X., Delgado, H., Tak, H., Jung, J., Shim, H., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T., Evans, N., Lee, K. A., & Yamagishi, J. (2024). ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *Proceedings of the ASVspoof Workshop 2024*.

Warren, K., Tucker, T., Crowder, A., Olszewski, D., Lu, A., Fedele, C., Pasternak, M., Layton, S., Butler, K., Gates, C., & Traynor, P. (2024). "Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (pp. 2696-2710). https://doi.org/10.1145/3658644.3670325