Can Spoofed Voices Compromise an ASR system?

Julia Roegner[1], *Simon Werner*[1], *Angelika Braun*[1]
[1]*Trier University, Germany*
s2juroeg@uni-trier.de

Given the availability and the quality of voice clones, they can be expected to show up in forensic casework on a regular basis (Kudera et al., 2024; San Segundo and Delgado, 2025).

This contribution addresses the question of whether a SOTA ASR system can deal with voice clones. To this end, one male native speaker of German provided the original speech samples. He was recorded on three consecutive days, producing six genuine samples. Depending on the recommendation of the manufacturer, one or two samples were used to clone his voice with two different models: a commercial one (ElevenLabs Inc., 2026) and an open-source one (Zyphra Technologies Inc., 2026). 30 clones per model were generated.

For testing, we compared the four remaining genuine samples to one cloned sample each. The first two testing modes (TMs) offered by Phonexia™ Voice Inspector (Phonexia, 2024) were used. In TM1, we applied one-to-one comparisons (4x60=240 genuine-cloned comparisons & 4x2=8 genuine-genuine comparisons), while in TM2, we combined the four genuine samples as material for the suspected speaker (1x60=60 genuine-cloned comparisons & 1x2 genuine-genuine comparisons).

The 300 comparisons (240+60) involving a cloned sample were judged to originate from different speakers 81% of the time (Fig. 1)[1]. The clones produced by the commercial model were more often classified correctly as coming from different speakers than those from the open-source model. The commercial model yielded correct rejection rates of 100% (TM1) and 50% (TM2), whereas the open-source model showed correct rejection rates of 90% (TM1) and 0% (TM2). The all-genuine comparisons had a correct rejection rate of 100% each.

We then looked at the probability ratings in terms of absolute log-likelihood ratios (LLRs, Fig. 2).

Between the clones, the absolute values were higher on average for the commercial model in TM1, and higher for the open-source model in TM2 despite that model's correct rejection rates of 0.0%. In TM1, the absolute LLR values of genuine sample lie in between those of the clones, and they show the highest values of the three models. Mann-Whitney U tests revealed significant differences of absolute values between correct and incorrect answers for samples of the same model (TM1 – open-source: $p<0.001$; TM2 – commercial: $p<0.05$). In both cases, the absolute values for incorrectly recognized samples are lower than the absolute values for correct answers. Further Mann-Whitney U tests showed a significant difference between clones depending on the model ($p<0.001$ in each testing mode).

In addition to the between-model variability, we observed some intra-model variability (see Fig. 2). In other words, two clones based on identical material may well produce categorically different results in terms of both correctness and range of LLR values. This as well as the influence of the testing method makes it difficult to draw general conclusions with respect to the effectiveness of cloned samples in deceiving ASR systems. The experiments show a relatively high probability for a correct result in one-to-one comparisons, but produced a large number of false positives in TM2. We would therefore emphasize the need to verify the authenticity of samples before applying ASR. To improve accuracy, the inclusion of synthesized speech in the training data of ASR might be useful. As a next step, we are planning to test the performance of human listeners as compared to ASR.

[1] Please note that a no-match answer was considered correct in the genuine-clone pairs, whereas a match was considered correct in the genuine-genuine pairs.

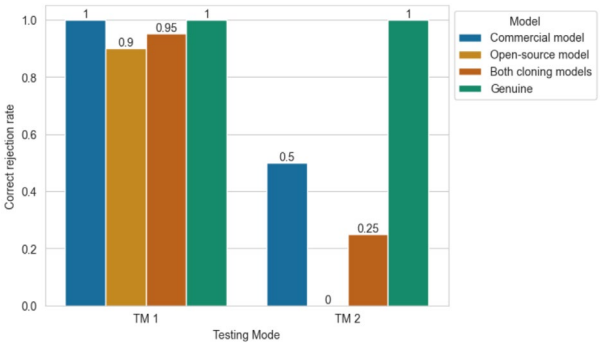
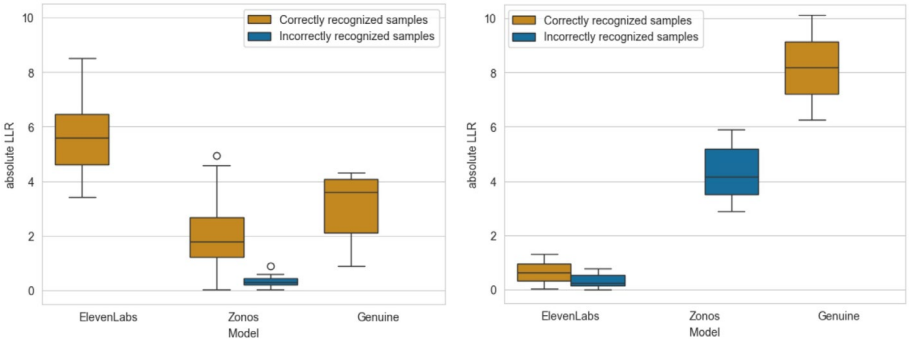


Figure 1. The correct rejection rates of samples by model, sorted by testing mode.



(a) TM1 (b) TM2

Figure 2. Boxplots for the absolute log-likelihood ratio (LLR) values throughout the two testing modes for each model, divided by correctly and incorrectly recognized samples.

References

Kudera, J., Coccia, M., Fadaeijouybari, S., Preidt, T., Ranjan, A., & Braun, A. (2024). Voice cloning and mismatch conditions in forensic automatic speaker recognition. In *International Conference on Speech and Computer* (pp. 171-184). Cham: Springer Nature Switzerland.

Phonexia (2024). *Voice inspector* (Version 5.2.0) [Computer software]. https://docs.phonexia.com/products/voice-inspector/download-voice-inspector-5-2

San Segundo, E., & Delgado, J. (2025). A qualitative systematic review of intra-speaker variation in the human voice. *Journal of Speech Sciences*, *14*(00), e025019.

Voice Cloning Models

ElevenLabs Inc. (2026): *ElevenCreative* (26 Jan Version) [AI-Platform to generate and alter audio and video]. https://elevenlabs.io/app/home

Zyphra Technologies Inc (2026): Zonos (26 Jan Version) [AI-model for voice cloning]. https://playground.zyphra.com/audio