Room Reverberation Effectively Masks Deepfake Traces

Sophie Hoppe, Anabell Hacker, Markus Brückl

Chair of Speech Communication, Technische Universität Berlin, Berlin, Germany
Email: {sophie.hoppe, anabell.hacker, markus.brueckl}@tu-berlin.de

Abstract

A study is presented on testing the performance of a publicly available AI tool (DEEPFAKE TOTAL) that is developed to analyse audio files to detect deepfakes.

The audio data in the test is built on a database of brand personality speech (BRANDDB), in which actors portray brand personalities in a way that is recognisable to listeners. Deepfakes were generated from these originals by using an established voice cloning AI (ELEVENLABS). Both the originals and the deepfakes were recorded a second time while being played back in an office space in order to add real-room reverberation. All recordings were analysed for fakeness by DEEPFAKE TOTAL.

The most crucial result is that deepfakes with room reverberation evaluate as not different from the originals. They even tend to be considered less fake than them, indicating that room reverberation effectively masks deepfake traces.

1 Motivation / Background

The rapid advancement of AI-based speech synthesis technologies has led to the proliferation of highly realistic audio deepfakes, which are becoming increasingly difficult to detect and pose significant threats to communication, media trust, and information security [1]. In light of these challenges, there is a pressing need for practical evaluations of deepfake detection technologies under realistic conditions. This study contributes to this discourse by testing a publicly available detection tool – DEEPFAKE TOTAL (DFT) [2, 3] – within a controlled experimental setup in which both synthetic and original speech data are acoustically altered to include room reverberation conditions.

The speech recordings used in this study are based on BRANDDB [4], a database developed in the context of brand personality research inspired by Aaker's [5] framework. The database contains recordings of professional actors impersonating the five brand personalities only by vocal means, while uttering the same linguistic content. Due to its controlled design and its availability, BRANDDB offers a robust foundation for replication and follow-up studies [4].

For the generation of deepfake audio samples, ELEVENLABS [6] is applied – a established AI-powered voice cloning system. ELEVENLABS is able to replicate voices with high expressiveness and naturalness. By using this system, the study ensures that the generated synthetic voices present a realistic challenge to detection tools, as they are barely distinguishable from recordings of genuine speech by human listeners.

The analysis of both the original and deepfake recordings is conducted using DFT [2, 3], a publicly accessible detection tool specifically designed to identify AI-generated audio content. The platform has gained recognition through media coverage and its use in real-world contexts such as security screening and media verification [7–9]. It employs a data-driven approach to detect typical artifacts of synthetic speech processing and provides non-experts with an initial assessment of potentially manipulated speech content [3]. Its accessibility and real-world applicability make it a suitable tool for this evaluation.

By combining these three components – a controlled, stylistically marked speech database (BRANDDB), a state-of-the-art deepfake synthesis system (ELEVENLABS), and a publicly available detection platform (DFT) –, the study suggests an experimental framework that is both scientifically grounded and practically relevant. It tests the effectiveness of common detection methods under room reverberation conditions and aims to contribute to the broader assessment of the robustness of current AI fake detection systems.

2 Data and Methods

All audio files, the creation of which is explained below, are published under the Creative Commons licence CC BY-NC-SA 4.0 and are available for download at: www.tu.berlin/en/kw/research/projects/branddb.

2.1 The original audio files

The original audio material (o) used in this study was sourced from BRANDDB [4], a speech corpus developed at Technische Universität Berlin to explore how brand personality is vocally expressed. The corpus is based on the established brand personality model by Aaker [5], consisting of the five dimensions *sincerity* (si), *excitement* (ex), *competence* (co), *sophistication* (so), and *ruggedness* (ru). Seven professional actors (three female (f), three male (m), one non-binary (d)) were selected in a casting, based on their demonstrated ability to convincingly and consistently vocalise all five brand personality dimensions.

The selected actors were recorded while uttering two German slogans – "Du bist unser Kunde, wir sind deine Marke" and "Sie sind unser Kunde, wir sind Ihre Marke" ("You are our customer, we are your brand" in both informal and formal form of address) – using only their voice and speaking style to convey a targeted brand personality dimension. The texts were designed to be semantically neutral across all dimensions while maintaining a marketing-related tone. Each actor performed the two slogans three times for each dimension.

Recordings were conducted in a low-reverberation, externally soundproofed speaker booth using a high-quality condenser microphone (AKG C414 XLII) and a digital audio interface (RME Fireface UC) at 48 kHz and 16-bit resolution. The recording setup deliberately avoided any additional signal processing. For each actor, 36 utterances were recorded: three per form of address and dimension, including the actor's 'neutral' personality. This resulted in a total of 252 recordings.

A perception study was conducted to validate the vocal representations of brand personality dimensions. A total of 79 listeners – mostly students of speech communication science or business administration – rated the recordings

using five sliders, one per dimension. The results confirmed statistically significant agreement among raters and showed a strong recognition of the intended brand personality dimensions across all speakers. In addition, the BRANDDB perception study identified for each speaker and each brand personality dimension a single utterance – regardless of whether it used the formal or informal address – that was rated as the most effective and recognisable expression of that specific dimension. These 7x5=35 serve as the 'best case' exemplars for analysis and synthesis in the study presented here.

2.2 Generation of the deepfake audios

2.2.1 Generation of the clone voices

To generate the deepfake audio material, the voice cloning service provided by ELEVENLABS is used, specifically the "Instant Voice Cloning" feature. For each of the seven original speakers from the BRANDDB corpus, a voice clone is created based on their full set of recorded utterances. The training input for each clone consists of all 36 utterances per speaker, which were concatenated into a single .flac audio file. One second of silence was inserted between each utterance to preserve temporal segmentation and ensure clean training transitions within the model.

During the cloning process, the default audio preprocessing option "Remove background noise from audio recordings" was explicitly disabled. This ensures that the subtle natural characteristics of the original voice, including breath sounds and micro-pauses, are preserved and included in the voice profile. These acoustic details contribute to the naturalness of the synthesised speech and help maintain fidelity to the vocal identity of the original speaker.

2.2.2 Cloning voice and speaking style

We use these 7 real-speaker simulating voice clones as voice models within the "Voice Changer" application in order to adapt not only the speakers' voices but also the speaking styles of their 'best-case' attempt. The voice model selected was "Eleven Multilingual v2", which supports nuanced prosodic and phonetic rendering across a broad range of languages. The voice cloning interface allows for fine-tuning along several dimensions. For this study, all parameters were kept at their default values.

This procedure results in one synthetic utterance per speaker per personality dimension, yielding five deepfake audio files per speaker.

The same configuration is consistently applied across all speakers to ensure comparability throughout the dataset. This approach not only maintains fidelity to the vocal profile of each speaker but also attempts to replicate their most convincing brand personality portrayal in a controlled synthetic setting.

Accordingly, this process produces 35 dimension-specific deepfake utterances (f).

2.3 Re-recording in a real room

To simulate the acoustic transformations that occur in every room, all audio samples – both original and synthetic – are re-recorded in a furnished office room. The recording space measures approximately 25 square metres, with a ceiling height of 3 meters, and contains typical furniture, contributing to natural sound reverberation and absorption; these conditions remained consistent across all recordings. The setup is designed to introduce real-room reverberation while maintaining speech intelligibility and consistent playback conditions. We assume that the general influence of room reverberation remains comparable regardless of the specific room used for re-recording. We chose an office environment because it represents an everyday setting and therefore closely reflects realistic application scenarios.

Audio playback is conducted using one active studio monitor speaker (Tannoy Reveal Active), driven by an audio interface (RME Fireface UC). The re-recording is captured using the same microphone as with the original recordings, positioned at a fixed listener-like distance (approximately 2 metres) from the speaker. The signal is digitised via an interface (Tascam US-144 MKII) with the same sampling frequency and resolution as the original. Playback and recording levels are adjusted to ensure a maximum sound pressure level without clipping.

The original and synthetic utterances are played back and recorded. Two distinct audio conditions are generated through this procedure. In the first condition, *room reverberation* (rr), we cut the sounds cleanly, so there are no pauses at the beginning and end of the utterance. In the second condition, *room reverberation with pause* (rrp), we used the same re-recording including a three-second pause before and after the utterance, producing a room intro and room coda.

From an auditory perspective, the resulting recordings exhibit noticeable room reverberation. Nevertheless, the spoken text, the intended brand personality dimension and the speaker identity remain intelligible and recognisable to human listeners.

2.4 Testing for deepfake

To assess the detectability of deepfake audio under room reverberation conditions as well as under control conditions, all six audio types (o, orr, orrp, f, frr, frrp) are submitted to DFT [2, 3]. The submission process is conducted manually: Each audio file is uploaded one-on-one to the web interface.

Upon analysis, the system returns a numerical fakeness score, expressed as a percentage with decimal precision (i.e., effectively 1001 tiers). This suggests a quasi-metrical scale measurement of a probability. Higher scores indicate a greater degree of suspected fakeness, whereas lower scores imply a stronger fit to naturally spoken human voice utterances.

2.5 Statistical methods

Thus, statistically speaking, the resulting raw data consist of six dependent samples, i.e., six repeated measurements, as defined by the six-tier factor *Audio type*: o, f, orr, frr, orrp, and frrp. Paired t-tests could be considered sufficient to test the (alternative) hypothesis that the audio types differ in DFT score. But since the data result from utterances that can be grouped by the gender of the speakers as well as the portrayed BP dimension, we also want to test simultaneously the influence of the factors *Brand personality dimension (BP dimension)* and speaker's *Gender* on the DFT score. We therefore consider the application of a three-way analysis of variance (ANOVA) – with a saturated model, including all interactions, type 3 square sums (cp. Table 1) – as the more suitable procedure. Since we do not analyse big data and an error would not cause enormous damage, we choose the common α error level of 5%. Since an ANOVA is an omnibus test, we make (undirected) pair-

$Effect$	DF_n	DF_d	SS_n	SS_d	F	p [%]	η^2_G [%]
Gender	2	20	334.11	24575.55	0.1359532	87.37	0.4
BP dimension	4	20	7062.50	24575.55	1.4368953	25.85	8.4
Audio type	5	100	160614.52	52699.96	60.95432	$7.938 \cdot 10^{-27}$	67.5
Gender : BP dim.	8	20	9638.45	24575.55	0.9804922	47.89	11.1
Gender : Audio type	10	100	12934.74	52699.96	2.4544111	1.15	14.3
BP dim. : Audio type	20	100	15719.21	52699.96	1.4913868	10.11	16.9
Gender : BP dim. : Audio type	40	100	15191.71	52699.96	0.7206698	87.80	16.4
$Corrections$					GGe	$p(GGe)$ [%]	
Audio type					0,521	$4.329 \cdot 10^{-14}$	
Gender : Audio type					0,521	4.323	

Table 1: Results from the ANOVA with the repeated measures factor *Audio type* and the two grouping factors *Gender* and *BP dimension* on DEEPFAKE TOTAL's scores.

wise comparisons of groupings from significant factors via Fisher's Least Significant Difference (LSD) test, again at an α level of 5%. For these computations and visualisations we use the function *ezANOVA* of the *ez* package [10] using R [11].

3 Results

The results of the 3-way-ANOVA with the repeated measures factor *Audio type* and the two grouping factors *Gender* and *BP dimension* can be found in Table 1: The factor *Audio type* and the interaction of *Gender* and *Audio type* show a significant effect on the values of the DFT scores. The Mauchly sphericity test indicates variance heterogeneity between the experimental groups (W=0.043; p=$4.514 \cdot 10^{-7}$). Therefore, the p-values of these factors in Table1 are corrected with the Greenhouse-Geisser (GGe) method – which still classifies the effect of the factor *Audio type* with p=$4.329 \cdot 10^{-14}$% as highly significant and its interaction with *Gender* as significant (p< 5%).

The effect sizes η^2_G can (at least roughly) be interpreted as explained variance, just like R^2 in regression analysis [12]. If one likes to interpret them in an absolute manner, Bakeman [13] can be followed: "Cohen (1988, pp. 413–414), who did not consider repeated measures designs explicitly, defined an η^2 [...] of .02 as small, one of .13 as medium, and one of .26 as large. It seems appropriate to apply the same guidelines to η^2_G as well." So, the effect of *Audio type* on the DFT scores can be regarded as very large and the effect of *Gender : Audio type* as medium.

Results form the pairwise comparisons (of group means) via Fisher's LSD can be found in Table 2 and Figure 1: Most obviously, it can be seen that the fake audio (f) is most of the time correctly classified by DFT with a score of nearly 100% for all genders. The clear difference in the mean scores of this audio type compared to all other audio types, of course, is greatly responsible for the large effect size of the factor *Audio type*.

The original audios (o) are judged less accurately, indicating a vulnerability of DFT to false positives. The largest difference from the expected fakeness of 0% is found in the female speakers' utterances with a mean value of 31.9%. Actually, this high mean is due to the fact that about one-third of the female audios is severely misjudged (e.g. f1_co=96.0%, f2_co=96.9%, f3_ex=95.6%, f3_si=99.5%), while two-thirds are rated quite properly. The resulting significant differences in the mean scores of females from males and from diverse account largely for the significant interaction of *Gender* and *Audio type*. This means that the speaker's gender significantly affects the fake recognition performance: The female speakers' originals are significantly judged more fake than those of other genders.

However, the most surprising and crucial result of our data pertains to the room reverberation fakes (the audio types frr and frrp): The mean scores on these room reverberation fakes cannot be significantly distinguished from those of the original audios by DFT. Moreover, they tend to be considered even more real than the originals.

The room reverberation originals (orr and orrp) do not contribute to making the results clearer overall: On the one hand, it seems plausible that the audio recordings altered by room reverberation are no longer classified as non-fake as clearly as the originals—after all, they were recorded a second time. On the other hand, it is noticeable that they are rated more fake than the corresponding room reverberation fakes.

Some systematicity may perhaps be gleaned from the comparison between the cleanly cut room reverberation sounds (rr) and those with 'room-filled' pauses before and after the actual utterance (rrp): Fakes and also originals tend to be rated as less fake when combined with a room intro and room coda, indicating a tendency of DFT to interpret any room reverberation as evidence for genuineness, falsely.

Gender	**Audio t.**	**N**	**Mean**	**LSD lo**	**LSD hi**
d	f	5	100.0	90.6	109.4
f	f	15	100.0	90.5	109.4
m	f	15	99.8	90.4	109.2
d	frr	5	12.4	3.0	21.8
f	frr	15	1.9	-7.5	11.3
m	frr	15	23.6	14.2	33.0
d	frrp	5	4.0	-5.4	13.4
f	frrp	15	2.1	-7.3	11.5
m	frrp	15	15.5	6.0	24.9
d	o	5	8.1	-1.3	17.5
f	o	15	31.9	22.5	41.3
m	o	15	1.4	-8.0	10.8
d	orr	5	42.1	32.7	51.5
f	orr	15	32.2	22.8	41.6
m	orr	15	35.1	25.7	44.5
d	orrp	5	32.7	23.3	42.1
f	orrp	15	24.1	14.7	33.5
m	orrp	15	33.1	23.7	42.5

Table 2: Means of DEEPFAKE TOTAL's scores (in %) grouped by *Audio type* and *Gender* and their lower and upper limits as defined by Fisher's LSD at an α level of 5%.

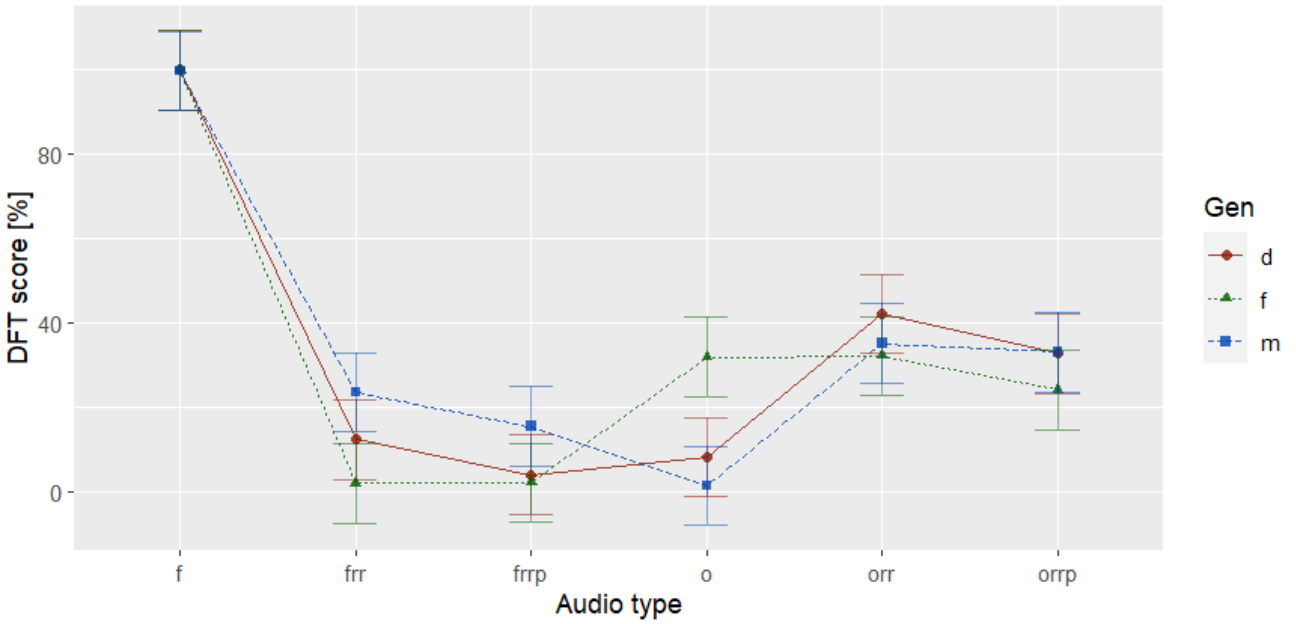

Figure 1: Means of DEEPFAKE TOTAL's scores grouped by *Audio type* and *Gender*. Error bars are defined by Fisher's LSD for an α level of 5% (LSD = 18.82%).

4 Discussion

For sure, our room recordings (orr, orrp, frr, and frrp) are real recordings after the second recording, not synthesised; thus, one could argue technically that they are not fake by definition and, therefore, quite appropriately rated by DFT. But then it can hardly be explained why the original utterances tend to be valued as more fake than the deepfakes after recorded a second time. But this is just a tendency, given our limited data. Aside from that, we do assume that DFT wants to detect whether the initial utterance is a deepfake, regardless of the final transmission path, which does not work satisfactorily at the moment, at least with our sounds, especially with the room reverberation fakes (frr, frrp). Both findings point to the fact that DFT falsely interprets the presence of room reverberation in a recording as a reliable evidence for the audio being no deepfake.

In a preliminary test, we also investigated whether additive noise has an influence on detectability; however, we did not find a significant effect. This result aligns with findings from other studies, which also report that model performance is largely unaffected by noise and that performance degradation in replay scenarios is more likely due to convolutional distortions than to noise itself [14, 15].

Given that DFT severely misjudges some original female audios as fakes and we, acoustic-phonetic speech scientists, are up to now unable to detect any systematicity in these deficits and considering all this together with the room reverberation deficits, we must preliminarily conclude that the features (i.e., the neural connections), on which DFT bases its assessments, are not valid in detecting deepfakes.

Note that we here only used ElevenLabs' "Instant Voice Cloning" and not "Professional Voice Cloning". It is possible that DFT could have even more problems with the more elaborated voice cloning version. That is to be tested in a follow-up.

Also, we are planning to expand the audio data in the fake detector test to 6 times the amount, using not only the 'best case' of each speaker, but all sounds of the database as speaking style models for generating optimal fakes, since we suggest that in addition to *Audio type* and *Gender*, the factor *BP dimension* should also become significant if more than our minimal set of audios is examined.

Of course, we have only tested one fake detector up to now. In order to generalise these results, we will have to get further fake detectors into the testing, but we suspect that other automatic deepfake detectors also have serious problems identifying fakes with room reverberation and that this is a problem that will be difficult to solve in a general manner.

5 Conclusion

In this study, we come to the conclusion that room reverberation effectively masks deepfake traces and that vocal gender significantly influences non-fake detection—at least in the automatic deepfake detection by DEEPFAKE TOTAL (DFT) by the Fraunhofer Institute for Applied and Integrated Security.

We suppose that these severe shortcomings arise from DFT using not appropriate or even misleading audio features in assessing (non-) fakeness, which themselves probably arise from a not sufficiently complex modelling, which again may be due to too few training sounds or at least too little variability in training sounds.

References

[1] Zentrum für vertrauenswürdige Künstliche Intelligenz, Berivan Köroglu, "KI-generierte Inhalte erkennen – Das Beispiel Deepfakes." https://www.zvki.de/ki-navigator/unsere-inhalte/ki-generierte-inhalte-erkennen-das-beispiel-deep, 2024. Accessed: 2025-05-20.

[2] Nicolas M. Müller, "DeepFake Total." https://deepfake-total.com/, 2025. Accessed: 2025-05-05.

[3] N. M. Müller, P. Sperl, and K. Böttinger, "Complex-valued

neural networks for voice anti-spoofing," in *Interspeech 2023*, pp. 3814–3818, 2023.

[4] M. Brückl, A. Hacker, N. Wünderlich, K. Talke, and D. Valeeva, "Eine Datenbank für Markensprechweise (BrandDB)," in *Proceedings of the 36th Conference on Electronic Speech Signal Processing (Konferenz Elektronische Sprachsignalverarbeitung) ESSV 2025*, Halle (Saale), Germany, Mar. 2025, pp. 130–137.

[5] J. L. Aaker, "Dimensions of brand personality," *Journal of Marketing Research*, vol. 34, no. 3, pp. 347–356, 1997.

[6] 11Labs Team, "ElevenLabs." https://elevenlabs.io/de, 2025. Accessed: 2025-04-21.

[7] J. O. Schneppat, "Deepfake total: Chancen, Risiken und Gegenmaßnahmen." https://gpt5.blog/deepfake-total-chancen-risiken-und-gegenmassnahmen/, 2024. Accessed: 2025-05-20.

[8] Deutsche Welle, Sparrow, "Faktencheck: Wie erkenne ich audio-deepfakes?." https://www.dw.com/de/faktencheck-wie-erkenne-ich-audio-deepfakes/a-69980269, 2024. Accessed: 2025-05-20.

[9] ZDF, "Bedrohung im Super-Wahljahr: Wie gefährlich sind Audio-Deepfakes?." https://www.zdf.de/nachrichten/wissen/audio-kuenstliche-intelligenz-deep-fakes-100.html, 2024. Accessed: 2025-05-20.

[10] M. A. Lawrence, *ez: Easy Analysis and Visualization of Factorial Experiments*, 2016. R package version 4.4-0.

[11] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[12] S. Olejnik and J. Algina, "Generalized eta and omega squared statistics: Measures of effect size for some common research designs," *Psychological Methods*, vol. 8, no. 4, pp. 434–447, 2003.

[13] R. Bakeman, "Recommended effect size statistics for repeated measures designs," *Behavior Research Methods*, vol. 37, no. 3, pp. 379–384, 2005.

[14] N. Müller, P. Kawa, W.-H. Choong, A. Stan, A. T. Bukkapatnam, K. Pizzi, A. Wagner, and P. Sperl, "Replay attacks against audio deepfake detection," 2025.

[15] H.-T. Luong, D.-T. Truong, K. A. Lee, and E. S. Chng, "Room impulse responses help attackers to evade deep fake detection," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 623–629, 2024.