IM RAUM DER TÄUSCHUNG – RAUMHALL ALS SCHWACHSTELLE AUTOMATISCHER DEEPFAKE-ERKENNUNG

Sophie Hoppe[1], *Anabell Hacker*[1,2], *Markus Brückl*[1]
[1]*Fachgebiet Kommunikationswissenschaft, Technische Universität Berlin, Germany*
[2]*Mobile Dialogsysteme, Otto-von-Guericke-Universität Magdeburg, Germany*
sophie.hoppe@tu-berlin.de

Kurzfassung: In dieser Studie wird die Robustheit eines frei zugänglichen Detektionssystems zur Erkennung von Audio-Deepfakes (DEEPFAKE TOTAL) getestet. Die verwendeten Äußerungen stammen aus der BRANDDB, in der sieben Schauspieler die fünf Markenpersönlichkeitsdimensionen ausschließlich durch Stimme und Sprechweise darstellen. Durch die Verwendung eines Voice-Cloning-Tools, sowie eines Voice-Conversion-Tools von ELEVENLABS, werden Audio-Deepfakes erzeugt, welche die Stimme und Sprechweise der Originaläußerungen synthetisieren. Sowohl die Originale als auch die Audio-Deepfakes werden zusätzlich in einem Büroraum erneut aufgenommen, um natürlichen Raumhall zu erzeugen.
Die Ergebnisse zeigen, dass durch Raumhall maskierte Audio-Deepfakes von DFT nicht mehr signifikant von Originalaudios unterschieden und teilweise sogar als natürlicher bewertet werden. Darüber hinaus weisen die Befunde auf genderspezifische Verzerrungen und den Einfluss der unterschiedlichen Sprechweise der Markenpersönlichkeitsdimensionen auf die Klassifikationsleistung hin. Alle Audiodateien und Ergebnisse sind unter `https://www.tu.berlin/kw/forschung/projekte` abzurufen.

1 Hintergrund

Moderne Verfahren der Sprachsynthese eröffnen heute ein breites Spektrum an Einsatzmöglichkeiten, etwa in der Medienproduktion, der Medizintechnik und der Unterhaltungsindustrie. In diesem Kontext kommen zunehmend Verfahren zum Einsatz, die sogenannte Audio-Deepfakes mit hoher Natürlichkeit erzeugen, welche für menschliche Hörer immer schwieriger von natürlich erzeugter Sprache zu unterscheiden sind.

Im Unterschied zu klassischen Ansätzen der Sprachsynthese wie Text-to-Speech (TTS), basieren moderne Sprachsyntheseverfahren auf der Verwendung eines zuvor erzeugten Stimmklons, der die Merkmale einer Zielstimme bestmöglich modelliert. Dieser Stimmklon kann dann genutzt werden, um einen beliebigen Text mit der geklonten Stimme zu sprechen [1, 2].

Darüber hinaus erlauben moderne Voice-Conversion-Systeme, wie etwa „Voice Changer“ [3] von ELEVENLABS, das gezielte Kopieren der Sprechweise (prosodische Eigenschaften wie Betonung, Rhythmus und Sprechfluss), welche bislang nur zusammen mit dem durch das Eingangsaudio vermittelten Textes kopierbar ist. Damit ist es aber möglich noch natürlicher klingende Audio-Deepfakes zu erzeugen, die noch näher die Äußerungen echter Menschen nachahmen können, weil sie eben nicht nur deren Stimme, sondern auch deren Sprechweise (zusammen mit dem Text) simulieren, wozu veraltete TTS-Systeme nicht in der Lage sind und wodurch sich selbstredend auch das Missbrauchspotenzial für Audio-Deepfakes erhöht [4].

Dementsprechend gewinnt die automatische Erkennung synthetischer Sprachsignale zunehmend an Bedeutung. Während aktuelle Detektionssysteme unter kontrollierten und idealisierten Bedingungen häufig hohe Erkennungsraten erzielen, rückt ihre Robustheit gegenüber

realistischen akustischen Bedingungen – etwa Raumhall oder variierenden Aufnahmesituationen wie dem Einsatz unterschiedlicher Mikrofone – erst in letzter Zeit verstärkt in den Fokus der Forschung [5].

Ein frei zugängliches System zur automatischen Erkennung von Audio-Deepfakes ist DEEPFAKE TOTAL (DFT) [6], welches in früheren Arbeiten bereits zur Analyse raumakustisch veränderter synthetischer Sprachsignale eingesetzt wurde [7]. Aufbauend auf dieser ersten explorativen Untersuchung, in der Schwachstellen bei der Detektion von durch Raumhall maskierten Audio-Deepfakes identifiziert werden konnten, wird dieser Ansatz in der vorliegenden Arbeit deutlich erweitert. Dazu wird der bislang nur teilweise genutzte BRANDDB-Korpus [8] in vollem Umfang ausgewertet.

Ziel dieser Untersuchung ist es, systematisch zu analysieren, wie sich realer Raumhall auf die Detektionsleistung von DFT auswirkt und inwiefern dieser Effekt durch das *Gender* der Sprecher:innen und die gezielte Modulation der Stimme und Sprechweise (über die Darstellung der verschiedenen Markenpersönlichkeitsdimensionen) beeinflusst wird. Durch die Nutzung des vollständigen Datensatzes soll überprüft werden, ob sich die von Hoppe et al. [7] zuvor beobachteten Effekte als konsistent erweisen und ob darüber hinaus weitere, bislang nicht erkannte Effekte wirksam werden.

2 Daten und Methode

2.1 Originalaudios

Die Originalaudios (o), die in dieser Untersuchung verwendet werden, stammen aus der BRANDDB [8], einem Korpus zur Realisierung von Markenpersönlichkeit nur über Stimme und Sprechweise. Das Korpus beinhaltet Darstellungen der fünf Markenpersönlichkeitsdimensionen (*BP dimension*) nach Aaker [9]: *Sincerity* (si), *Excitement* (ex), *Competence* (co), *Sophistication* (so) und *Ruggedness* (ru).

Für die Aufnahmen wurden sieben professionelle Sprecher:innen (drei weibliche (w), drei männliche (m) und eine divers-geschlechtliche (d) Person) ausgewählt. Die Sprecher:innen äußerten zwei deutschsprachige Werbeslogans – „Du bist unser Kunde, wir sind deine Marke“ (informell) und „Sie sind unser Kunde, wir sind Ihre Marke“ (formell). Die Slogans wurden so gestaltet, dass sie bezüglich der Markenpersönlichkeitsdimension neutral, aber dennoch inhaltlich marketingbezogen sind. Zusätzlich wurde von jeder Sprecher:in ein Set aus sechs Äußerungen aufgenommen, die in keiner Markenpersönlichkeit, sondern in der Persönlichkeit der Sprecher:in dargeboten wurden. Jede Sprecher:in äußerte jede Kombination aus Slogan (formell/informell) und Markenpersönlichkeitsdimension jeweils dreimal, sodass (zusammen mit den sechs neutralen Äußerungen) 36 Äußerungen pro Sprecher:in entstanden.

Die Aufnahmen erfolgten dabei in einer schallarmen, akustisch optimierten Sprecherkabine, mit einem Großmembran-Kondensator-Mikrofon (AKG C414 XLII) und einem digitalen Audiointerface (RME Fireface UC) mit einer Samplingrate von 48 kHz und einer Auflösung von 16-Bit, wobei jegliche weitere Signalverarbeitung vermieden wurde.

Aufgrund dieser Konzeption bietet die BRANDDB zwei zentrale Vorteile für die vorliegende Untersuchung: Erstens ermöglicht sie durch die gezielte Umsetzung unterschiedlicher Markenpersönlichkeitsdimensionen eine strukturierte Variation von Stimme und Sprechweise. Zweitens bleibt der Text der Äußerungen in den Slogans über alle Bedingungen und Sprecher:innen hinweg konstant, sodass mögliche Unterschiede bei der Detektion durch DFT gezielt auf Merkmale der Stimme und Sprechweise zurückgeführt werden können.

Für die vorliegende Untersuchung wurden alle im Korpus enthaltenen dimensionsspezifischen Aufnahmen berücksichtigt, um eine breitere Datenbasis zu schaffen und interne Variabilität der synthetisierten Stimuli zu ermöglichen. Die neutralen Äußerungen, die keiner expliziten

Dimension zugeordnet sind, wurden nicht mit einbezogen. Der Datensatz der Originalaudios (o) umfasst so insgesamt $N = 210$ Audiodateien.

2.2 Erzeugung der Audio-Deepfakes

2.2.1 Generierung der Stimmklone

Zur Erstellung der Stimmklone wurde das Voice-Cloning-Tool von ELEVENLABS genutzt, konkret die Funktion „Instant Voice Cloning“ [10]. Von jeder/jedem der sieben Sprecher:innen wurde dabei ein individueller Stimmklon erzeugt. Als Trainingsgrundlage diente jeweils das vollständige Set der verfügbaren Äußerungen pro Sprecher:in (insgesamt 36 Einzelaufnahmen).

Im Rahmen der automatischen Vorverarbeitung des Audiomaterials wurde die Standardoption „Remove background noise from audio recordings“ bewusst deaktiviert. Dadurch sollten feine sprecher:innen-spezifische Merkmale, wie Atemgeräusche, Mikro-Pausen und andere Merkmale natürlicher Äußerungen erhalten bleiben, welche sowohl zur natürlicheren Stimmqualität der Stimmklone beitragen, als auch die individuellen stimmlichen Identitätsmerkmale der Originalsprecher:innen möglichst authentisch abbilden.

2.2.2 Klonen der Stimme und Sprechweise

Um nicht nur die Stimmen der Sprecher:innen, sondern auch deren Sprechweisen zu simulieren, wird das „Voice“-Conversion-Tool „Voice Changer“ [3] von ELEVENLABS, unter der Verwendung des „eleven_multilingual_sts_v2“ Modells [11], verwendet. Dieses Tool erlaubt eine Feinabstimmung über verschiedene Syntheseargumente (stability: 1.0, similarity_boost: 1.0, style: 0.0, use_speaker_boost: True, remove_background_noise: False), welche über alle Sprecher:innen hinweg konsistent beibehalten werden, um die Vergleichbarkeit des gesamten Datensatzes zu gewährleisten.

Auf diese Weise kann mit dem „Voice Changer“ für jede Äußerung der Sprecher:innen ein Audio-Deepfake erzeugt werden, bei dem durch das „Voice“-Conversion-Verfahren nicht nur der textliche Inhalt, sondern auch die Sprechweise des jeweiligen Eingangsmaterials übernommen werden. Entsprechend entsteht für jedes der 210 Originalaudios (o) ein synthetisches Audio-Deepfake (f) als Gegenstück.

2.3 Raumakustische Neuaufnahme

Um die akustische Transformation zu erzeugen, die bei der Realisation von gesprochener Sprache in realen Räumen auftritt, wurden alle Audiodateien – sowohl die Originalaudios (o), als auch die synthetisch generierten Audio-Deepfakes (f) – in einem möblierten Büroraum erneut aufgenommen. Der Raum maß etwa 25 m^2 bei einer Deckenhöhe von 3 Metern und war mit typischen Möbelstücken ausgestattet, wodurch sowohl natürliche Reflexionen, als auch Schallabsorption erreicht wurden. Ziel dieser Maßnahme war es, natürlichen Raumklang zu generieren, ohne die Sprachverständlichkeit oder die Vergleichbarkeit der Aufnahmen wesentlich zu beeinträchtigen.

Die Wiedergabe der Audiodateien erfolgte über einen aktiven Studiomonitor (Tannoy Reveal Active), der über ein Audiointerface (Tascam US-144 MKII) angesteuert wurde. Die Neuaufnahmen wurden anschließend mit demselben Mikrofon (AKG C414 XLII) und unter denselben technischen Bedingungen erstellt, wie die ursprünglichen Aufnahmen des BRANDDB-Korpus, einschließlich der Abtastrate von 48 kHz und einer 16-Bit-Auflösung über das digitale Audiointerface RME Fireface UC. Das Mikrofon wurde dabei in einem Abstand von etwa 2 Metern, der dem einer normalen Konversation entspricht, zur Schallquelle fixiert. Die Wiedergabe- und Aufnahmepegel wurden so eingestellt, dass ein hoher Signalpegel erreicht wurde, ohne

Clipping zu verursachen.

Die so entstandenen Audios mit Raumhall werden in zwei Varianten erstellt: mit 3-sekündigem Raum-Intro und einer eben so langen -Coda (room reverberation with pause: rrp), also drei Sekunden, in denen nur der Raum hörbar ist, vor und nach der Äußerung, sowie ohne diese „Pausen“, also direkt vor dem ersten und nach dem letzten Phonem geschnitten (room reverberation: rr). Insgesamt entstanden so $N = 840$ raumakustisch veränderte Audios. Zusammen mit den Originalaudios (o) und den Audio-Deepfakes (f) umfasst der Datensatz somit $N = 1260$ Audios, die durch die drei Faktoren *Audio type* (o, orr, orrp, f, frr, frrp), *Gender* (w, m, d) und *BP dimension* (co, ex, ru, si, so) gruppiert werden können.

2.4 Testverfahren und statistische Analyse

Um die Erkennbarkeit der Audio-Deepfakes zu testen, wurden alle 1260 Äußerungen über das Webinterface von DFT hochgeladen und analysiert [6].

2.4.1 Bewertungsmetrik von DFT

Nach der Analyse gibt DFT für jede Audiodatei einen numerischen "Fake-O-Meter-Score"(DFT-Score) aus, der die prozentuale Einschätzung der Wahrscheinlichkeit abbildet, dass es sich um ein synthetisches Sprachsignal handelt. Der Score liegt auf einer Skala von $0,0$ % bis $100,0$ %, wobei niedrigere Werte eine höhere Übereinstimmung mit natürlicher Sprache anzeigen und höhere Werte auf eine synthetische Generierung hindeuten.

Der ausgegebene Wert entspricht dem arithmetischen Mittel der während der Analyse berechneten Einzelscores über zeitliche Analysefenster (der Dauer 3,23 Sekunden) [12]. Für die statistische Auswertung wurde in allen Fällen der von DFT bereitgestellte, systemseitig gerundete Mittelwert verwendet. Dieses Vorgehen gewährleistet eine konsistente und vergleichbare Bewertung über alle Äußerungen hinweg, unabhängig von deren Länge oder Anzahl der analysierter Zeitfenster.

2.4.2 Statistische Methoden

Ziel der statistischen Auswertung ist es, systematische Unterschiede in der Detektion von Audio-Deepfakes durch das System DFT aufzudecken. Die Rohdaten bestehen aus sechs abhängigen Stichproben, entsprechend den sechs Ausprägungen des Faktors *Audio type* (o, orr, orrp, f, frr, frrp), die als Messwiederholung statistisch umgesetzt werden.

Da die Audiodaten zusätzlich nach dem *Gender* der Sprecher:innen und den Markenpersönlichkeitsdimension (*BP dimension*) gruppiert werden können, wird zur gleichzeitigen Analyse aller Einflussgrößen eine dreifaktorielle Varianzanalyse (ANOVA) verwendet. Die ANOVA wird als saturiertes Modell mit Typ-III-Quadratsummen berechnet und umfasst somit alle Haupteffekte sowie Interaktionen.

Für alle Analysen wird ein Signifikanzniveau von α=0,05 angesetzt. Da die ANOVA ein Omnibustest ist, werden signifikante Effekte durch paarweise Post-hoc-Vergleiche mittels Fisher's Least Significant Difference (FLSD) weiter differenziert (ebenfalls α=0,05). Die statistischen Auswertungen und Visualisierungen erfolgen mit der Funktion *ezANOVA* aus dem *ez*-Paket [13] unter Verwendung der Statistiksoftware R [14].

3 Ergebnisse

Die ANOVA-Ergebnisse sind in Tabelle 1 dargestellt. Der Mauchly-Test auf Sphärizität weist auf Varianzheterogenität zwischen den abhängigen Experimentalgruppen hin ($W = 0{,}038$, $p = 3{,}66 \cdot 10^{-125}$), weshalb die Greenhouse-Geisser-Korrektur (GGe) verwendet wird. Auch nach dieser Korrektur zeigt der Faktor *Audio type* einen hochsignifikanten Effekt auf die DFT-Scores

($p = 2{,}21 \cdot 10^{-164}$). Darüber hinaus sind sowohl die Interaktion *Gender* × *Audio type* ($p = 0{,}107\,\%$), als auch die dreifache Interaktion *Gender* × *BP dimension* × *Audio type* signifikant ($p = 0{,}973\,\%$).

Effect	DF_n	DF_d	SS_n	SS_d	F	p [%]	η^2_G [%]
Gender	2	195	1910,658	45294,66	4,112828	1,78	0,85
BP dimension	4	195	1518,757	45294,66	1,634617	16,71	0,68
Audio type	5	975	1225278,082	176958,6	1350,198472	$0{,}00{\cdot}10^{+00}$	84,65
Gen. : BP d.	8	195	4474,377	45294,66	2,407855	1,69	1,97
Gen. : Audio t.	10	975	8765,044	176958,6	4,829332	$8{,}23{\cdot}10^{-05}$	3,79
BP d. : Audio t.	20	975	3979,596	176958,6	1,096332	34,69	1,76
Gen. : BP d.: Audio t.	40	975	15169,346	176958,6	2,089488	0,01	6,39
Corrections					*GGe*	$p(GGe)$ [%]	
Audio t.					0,378	$2{,}21{\cdot}10^{-164}$	
Gen. : Audio t.					0,378	0,107	
Gen. : BP dim.: Audio t.					0,378	0,973	

Tabelle 1 – Ergebnisse der ANOVA mit dem Faktor *Audio type* (Messwiederholung) und den beiden Gruppierungsfaktoren *Gender* und *BP dimension* auf den DFT-Scores.

Die Effektstärken η^2_G lassen sich näherungsweise als erklärte Varianz interpretieren und sind mit dem Determinationskoeffizienten R^2 aus Regressionsanalysen vergleichbar [15]. Zur groben inhaltlichen Einordnung können dabei die von Bakeman vorgeschlagenen Richtwerte herangezogen werden – die auf Cohen (1988, S. 413-414) zurückgehen und nach Bakeman auch sinnvoll auf η^2_G übertragbar sind – wonach ein η^2 ab 0,02 als kleiner, ab 0,13 als mittlerer und ab 0,26 als großer Effekt gilt [16]. Der Effekt des Faktors *Audio type* mit $\eta^2_G = 84{,}65\,\%$ ist demnach als sehr groß einzustufen, wohingegen der Effekt der Interaktionen *Gender* × *Audio type* ($\eta^2_G = 3{,}79\,\%$) und *Gender* × *BP dimension* × *Audio type* ($\eta^2_G = 6{,}39\,\%$) als klein zu bewerten sind.

Die paarweisen Vergleiche mittels Fisher's Least Significant Difference (FLSD) Abbildung 1 zeigen deutlich, dass DFT Audio-Deepfakes ohne Raumhall (f) nahezu fehlerfrei erkennt ($\bar{x}_{DFT-Score_f} = 99{,}02\,\%$). Ist Raumhall im Signal vorhanden, nimmt die Erkennungsleistung des Systems erheblich ab: Audio-Deepfakes mit Raumhall (frr) werden im Mittel als natürlicher bewertet ($\bar{x}_{DFT-Score_{frr}} = 2{,}06\,\%$), als die Originalaudios (o) ($\bar{x}_{DFT-Score_o} = 11{,}11\,\%$). Enthalten die Audio-Deepfakes zusätzlich Raumhall-Intro und -Coda (frrp) verstärkt sich dieser Effekt ($\bar{x}_{DFT-Score_{frrp}} = 1{,}69\,\%$). Zudem werden die Originalaudios mit Raumhall (orr, orrp) durch DFT als weniger natürlich bewertet als ihre synthetischen Gegenstücke in den selben *Audio type* Bedingungen ($\bar{x}_{DFT-Score_{orr}} = 3{,}70\,\%$, $\bar{x}_{DFT-Score_{orrp}} = 2{,}19\,\%$). Sowohl Originale als auch Audio-Deepfakes werden in Anwesenheit eines Raum-Intros und einer Raum-Coda als weniger synthetisch bewertet. Dies deutet darauf hin, dass DFT Raumhall fälschlich als Indikator für Natürlichkeit interpretiert.

Originalaudios (o) von weiblichen Sprecherinnen (w) werden zudem häufiger fälschlich als Fakes klassifiziert ($\bar{x}_{DFT-Score_{o,w}} = 19{,}24\,\%$), als die anderer *Gender* ($\bar{x}_{DFT-Score_{o,d}} = 8{,}41\,\%$, $\bar{x}_{DFT-Score_{o,m}} = 5{,}69\,\%$), was auf eine genderspezifische Verzerrung im zugrundeliegenden Modell des Systems hinweist.

Die auffälligste Ausnahme von dieser Systematik ist bei der *BP dimension* Ruggedness (ru) zu beobachten, wodurch größtenteils auch die Dreifachinteraktion begründet ist: Hier werden die originalen Äußerungen (o) der diversen Person (d) signifikant als noch synthetischer, als diejenigen der Männer (m) und diese wiederum signifikant synthetischer, als die der Frauen (w) durch DFT beurteilt.

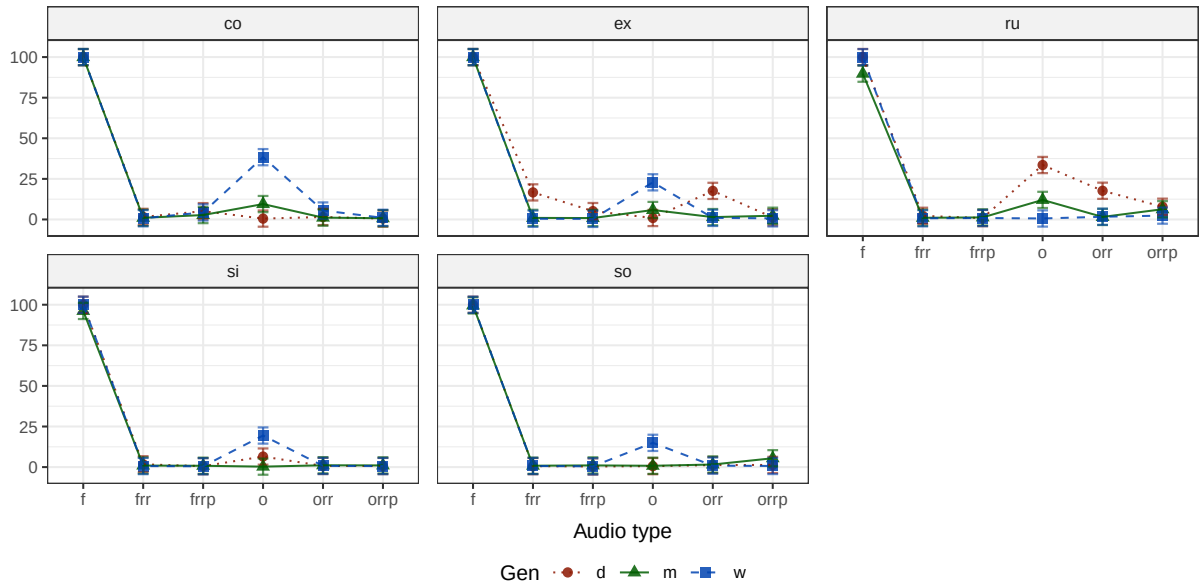

Abbildung 1 – Mittelwerte der DFT-Scores, getrennt nach *BP dimension* und gruppiert nach *Audio type* und *Gender*. Die Fehlerbalken sind durch Fishers LSD für einen α-Niveau von 5 % (FLSD = 4,58 %) bestimmt. Die genauen Werte dieser Abbildung lassen sich unter `https://www.tu.berlin/kw/forschung/projekte` abrufen.

4 Diskussion

Die vorliegenden Ergebnisse bestätigen und erweitern frühere Befunde zur Anfälligkeit automatischer Audio-Deepfake-Erkennung gegenüber realer Raumakustik [5, 7]. Während DFT unter kontrollierten, hallfreien Bedingungen Audio-Deepfakes zuverlässig identifiziert, zeigt sich unter realistischen Raumakustik-Bedingungen ein drastischer Leistungsabfall. Insbesondere können Audio-Deepfakes mit Raumhall nicht mehr von natürlichen Sprachaufnahmen unterschieden werden und werden teilweise sogar als weniger synthetisch bewertet als Originalaudios. Damit wird deutlich, dass Raumhall Artefakte synthetischer Erzeugung effektiv maskiert und eine zentrale Schwachstelle des hier verwendeten Systems offenlegt.

Im Vergleich zur vorangegangenen Studie auf Basis eines reduzierten Datensatzes erlaubt die Nutzung des vollständigen BRANDDB-Korpus eine deutlich robustere statistische Absicherung dieser Effekte. Die Ergebnisse zeigen, dass die beobachteten Fehlklassifikationen keine Einzelfälle oder Ausreißer darstellen, sondern sich systematisch über Sprecher:innen, Markenpersönlichkeitsdimensionen und akustische Varianten hinweg reproduzieren lassen. Damit verschiebt sich die Interpretation von einer explorativen Beobachtung hin zu einem belastbaren Befund über die Grenzen aktueller Detektionsansätze.

Besonders aufschlussreich ist der Befund, dass Raum-Intro und -Coda die Fehlklassifikation zusätzlich verstärken. Dies deutet darauf hin, dass DFT die akustische Präsenz des Raumes als Indikator für Natürlichkeit missinterpretiert. Ein solches Entscheidungsverhalten ist problematisch, da es reale Übertragungs- und Wiedergabeszenarien systematisch bevorzugt, unabhängig davon, ob das zugrundeliegende Signal ursprünglich synthetisch erzeugt wurde.

Die signifikante Interaktion zwischen *Gender* und *Audio type* bestätigt zudem frühere Hinweise auf genderspezifische Verzerrungen in der Klassifikationsleistung. Insbesondere Originalaudios weiblicher Sprecherinnen werden häufiger fälschlich als Audio-Deepfakes bewertet. Die vorliegende Untersuchung zeigt, dass dieser Effekt auch bei größerer Datenbasis bestehen bleibt und somit nicht auf zufällige Eigenschaften einzelner Äußerungen zurückzuführen ist. Welche Merkmale hierbei ausschlaggebend sind, lässt sich auf Basis der vorliegenden Analyse jedoch nicht eindeutig bestimmen. Allerdings deuten die Ergebnisse auf eine nicht realistische repräsentative Verteilung bestimmter phonatorischer oder prosodischer Eigenschaften im Mo-

dell hin, was zu Überanpassung an bestimmte Artefakte oder Muster führen muss.

Über die bisherigen Arbeiten hinaus zeigt sich erstmals eine signifikante dreifache Interaktion zwischen *Gender*, *BP dimension* und *Audio type*. Auch wenn die Stärke dieses Effekts vergleichsweise klein ist, weist er darauf hin, dass die Detektionsleistung von DFT nicht nur von isolierten Faktoren abhängt, sondern durch komplexe Kombinationen stimmlicher und/oder prosodischer Eigenschaften beeinflusst wird. Dies unterstreicht die Notwendigkeit, stimmliche Vielfalt und kontextuelle Variabilität stärker in die Entwicklung und Evaluation von Deepfake-Detektionssystemen einzubeziehen.

Insgesamt legen die Ergebnisse nahe, dass die von DFT genutzten Merkmalsrepräsentationen unter realen akustischen Bedingungen nicht zur Erkennung von Fakes eingesetzt werden können. Die beobachteten Fehlklassifikationen sprechen dafür, dass Merkmale der Raumakustik als Proxy für Natürlichkeit herangezogen werden, obwohl sie tatsächlich keinen verlässlichen Hinweis auf die synthetische oder natürliche Herkunft eines Signals liefern. Diese Problematik ist nicht auf DFT beschränkt, sondern betrifft auch andere datengetriebene Detektionssysteme, die überwiegend auf nicht ausreichend diversifizierten Trainingsdaten basieren [5].

5 Fazit

Die vorliegende Arbeit untersucht die Robustheit automatischer Audio-Deepfake-Erkennung gegenüber realistischen raumakustischen Bedingungen am Beispiel des Detektionssystems DEEPFAKE TOTAL (DFT). Auf Basis des vollständigen BRANDDB-Korpus konnte gezeigt werden, dass DFT unter hallfreien Bedingungen synthetische Sprachsignale zwar zuverlässig erkennt, jedoch bereits durch natürlichen Raumhall erheblich in seiner Erkennungsleistung beeinträchtigt wird. Audio-Deepfakes mit Raumhall werden systematisch als natürlicher klassifiziert als Audio-Deepfakes ohne Raumhall und sind momentan von DFT nicht mehr von Originalaufnahmen zu unterscheiden.

Im Vergleich zu einer vorangegangenen explorativen Studie erlaubt die Verwendung des vollständigen Datensatzes eine deutlich bessere statistische Absicherung dieser Befunde. Neben dem dominanten Einfluss des *Audio type*s konnten erstmals auch signifikante Interaktionen zwischen *Gender*, *BP dimension* und *Audio type* nachgewiesen werden. Diese Ergebnisse verdeutlichen, dass die Performanz von Detektionssystemen nicht allein von der synthetischen Herkunft eines Signals abhängt, sondern auch maßgeblich durch Variabilität der Stimme und Sprechweise und (raum-) akustischen Kontext beeinflusst wird.

Die Befunde legen nahe, dass Raumhall von DFT als Indikator für Natürlichkeit fehlinterpretiert wird und somit eine strukturelle Schwachstelle aktueller Detektionsansätze darstellt. Für zukünftige Forschung ergibt sich daraus die Notwendigkeit, raumakustische Variabilität systematisch in Trainings- und Evaluationsdaten einzubeziehen sowie stimmliche Diversität stärker zu berücksichtigen. Nur durch realitätsnahe Trainingsdatensätze und interdisziplinäre Perspektiven lassen sich robuste, faire und verlässliche Verfahren zur Erkennung von Audio-Deepfakes entwickeln.

Literatur

[1] BSI: *Deepfakes - gefahren und gegenmaßnahmen*. 2025. URL `https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html`. Zugriff: 2025-07-04.

[2] ELEVENLABS TEAM: *Instant voice cloning*. 2025. URL `https://elevenlabs.io/`

docs/creative-platform/voices/voice-cloning/instant-voice-cloning. Zugriff: 2026-01-24.

[3] ELEVENLABS TEAM: *Voice changer*. 2026. URL https://elevenlabs.io/docs/creative-platform/playground/voice-changer. Zugriff: 2026-01-24.

[4] KÖROGLU, B.: *KI-generierte Inhalte erkennen – Das Beispiel Deepfakes*. *Zentrum für vertrauenswürdige Künstliche Intelligenz*, 2024. URL https://www.zvki.de/ki-navigator/unsere-inhalte/ki-generierte-inhalte-erkennen-das-beispiel-deep. Zugriff: 2025-07-01.

[5] MÜLLER, N., P. KAWA, W.-H. CHOONG, A. STAN, A. T. BUKKAPATNAM, K. PIZZI, A. WAGNER, und P. SPERL: *Replay attacks against audio deepfake detection*. *Interspeech 2025*, 2025. URL https://arxiv.org/abs/2505.14862. Online-Vorveröffentlichung, 2505.14862.

[6] MÜLLER, N., P. SPERL, und K. BÖTTINGER: *Complex-valued neural networks for voice anti-spoofing*. In *Interspeech 2023*, S. 3814–3818. 2023. doi:10.21437/Interspeech.2023-901.

[7] HOPPE, S., A. HACKER, und M. BRUECKL: *Room reverberation effectively masks deepfake traces*. In *Speech Communication; 16th ITG Conference*, S. 106–110. 2025. URL https://ieeexplore.ieee.org/document/11264404.

[8] BRÜCKL, M., A. HACKER, N. WÜNDERLICH, K. TALKE, und D. VALEEVA: *Eine Datenbank für Markensprechweise (BrandDB)*. In *Proceedings of the 36th Conference on Electronic Speech Signal Processing (Konferenz Elektronische Sprachsignalverarbeitung) ESSV 2025*, S. 130–137. Halle (Saale), Germany, 2025.

[9] AAKER, J. L.: *Dimensions of brand personality*. *Journal of Marketing Research*, 34(3), S. 347–356, 1997.

[10] ELEVENLABS TEAM: *Voice cloning*. 2025. URL https://elevenlabs.io/docs/creative-platform/voices/voice-cloning. Zugriff: 2026-01-24.

[11] ELEVENLABS TEAM: *ElevenLabs Documentation: Models overview*. ElevenLabs, New York, USA, 2025. URL https://elevenlabs.io/docs/models#models-overview. Zugriff: 2025-07-11.

[12] MÜLLER, N.: *DeepFake Total*. 2025. URL https://deepfake-total.com/. Zugriff: 2025-07-05.

[13] LAWRENCE, M. A.: *ez: Easy Analysis and Visualization of Factorial Experiments*, 2016. doi:10.32614/CRAN.package.ez. URL https://CRAN.R-project.org/package=ez. R package version 4.4-0.

[14] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[15] OLEJNIK, S. und J. ALGINA: *Generalized eta and omega squared statistics: Measures of effect size for some common research designs*. *Psychological Methods*, 8(4), S. 434–447, 2003.

[16] BAKEMAN, R.: *Recommended effect size statistics for repeated measures designs*. *Behavior Research Methods*, 37(3), S. 379–384, 2005.