Vishing: detecting social engineering in spoken communication — a first survey & urgent roadmap to address an emerging societal challenge

Andreas Triantafyllopoulos, Anika A. Spiesberger, Iosif Tsangko, Xin Jing, Verena Distler, Felix Dietz, Florian Alt, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Triantafyllopoulos, Andreas, Anika A. Spiesberger, Iosif Tsangko, Xin Jing, Verena Distler, Felix Dietz, Florian Alt, and Björn W. Schuller. 2025. "Vishing: detecting social engineering in spoken communication — a first survey & urgent roadmap to address an emerging societal challenge." Computer Speech & Language 94: 101802. https://doi.org/10.1016/j.csl.2025.101802.

Computer Speech and Language 94 (2025) 101802

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Vishing: Detecting social engineering in spoken communication — A first survey & urgent roadmap to address an emerging societal challenge

Andreas Triantafyllopoulos [a,c,e,*], Anika A. Spiesberger [a,c,e], Iosif Tsangko [a,c,e], Xin Jing [a,c,e], Verena Distler [b], Felix Dietz [b], Florian Alt [b], Björn W. Schuller [a,c,d,e,f]

[a] CHI – Chair of Health Informatics, TUM University Hospital, Munich, Germany
[b] CODE – Research Institute for Cyber Defense, University of the Bundeswehr Munich, Munich, Germany
[c] EIHW – Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany
[d] GLAM – Group on Language, Audio, & Music, Imperial College London, London, United Kingdom
[e] MCML – Munich Center for Machine Learning, Technical University of Munich, Munich, Germany
[f] MDSI – Munich Data Science Institute, Technical University of Munich, Munich, Germany

ARTICLE INFO

Keywords:
Vishing
Social engineering
Human–computer interaction
Computational paralinguistics

ABSTRACT

Vishing – the use of voice calls for phishing – is a form of Social Engineering (SE) attacks. The latter have become a pervasive challenge in modern societies, with over 300,000 yearly victims in the US alone. An increasing number of those attacks is conducted via voice communication, be it through machine-generated 'robocalls' or human actors. The goals of 'social engineers' can be manifold, from outright fraud to more subtle forms of persuasion. Accordingly, social engineers adopt multi-faceted strategies for voice-based attacks, utilising a variety of 'tricks' to exert influence and achieve their goals. Importantly, while organisations have set in place a series of guardrails against other types of SE attacks, voice calls still remain 'open ground' for potential bad actors. In the present contribution, we provide an overview of the existing speech technology subfields that need to coalesce into a protective net against one of the major challenges to societies worldwide. Given the dearth of speech science and technology works targeting this issue, we have opted for a narrative review that bridges the gap between the existing psychological literature on the topic and research that has been pursued in parallel by the speech community on some of the constituent constructs. Our review reveals that very little literature exists on addressing this very important topic from a speech technology perspective, an omission further exacerbated by the lack of available data. Thus, our main goal is to highlight this gap and sketch out a roadmap to mitigate it, beginning with the psychological underpinnings of vishing, which primarily include deception and persuasion strategies, continuing with the speech-based approaches that can be used to detect those, as well as the generation and detection of AI-based vishing attempts, and close with a discussion of ethical and legal considerations.

1. Introduction

Social engineering (SE) has become one of the most prominent attack vectors in cybercrime. By manipulating their victims, cybercriminals seek to make users reveal credentials, issue payments, or disclose confidential information. A recent report by the

* Corresponding author.
E-mail address: andreas.triantafyllopoulos@tum.de (A. Triantafyllopoulos).

https://doi.org/10.1016/j.csl.2025.101802
Received 5 August 2024; Received in revised form 16 December 2024; Accepted 20 March 2025
Available online 15 April 2025

FBI found that 323,972 individuals fell prey to SE attacks in 2021 alone, a stark increase from previous years.[1] These attacks resulted in an estimated cost of $6.9 billion, making SE a pressing security issue with important financial (and psychological) ramifications. Crucially, humans play an integral role in 80–90% of successful cyberattacks, as making users disclose sensitive information is often much easier than circumventing computer security protocols. Accordingly, SE attacks exploit different media to approach their human targets, with the most widely-known form of attack being emails – an activity known as *phishing*. Another increasingly popular form of SE attacks is *vishing* – using voice for phishing.

During a vishing attack, attackers call their (prospective) victims and talk them into performing several actions (Krombholz et al., 2015). Common examples are fake calls from IT departments, asking users to provide their login username and password to mitigate an imminent cyber attack; calls in which criminals pretend to be from the police; or cases in which attackers pretend to be a relative in dire need of some money (potentially deep faking a known voice).

Vishing is becoming an increasingly popular form of social engineering because attackers can adapt flexibly to their victim's responses, and the victim has no time to step back to consider the attack due to the immediate nature of the communication. The perceived urgency might thus be even higher than in the case of phishing. Beyond raising awareness, few effective means of protection exist against such attacks today.

We expect the situation to get even more dramatic. The rise of ubiquitous text-to-speech (TTS) algorithms and – particularly – affective TTS systems (Triantafyllopoulos et al., 2023; Triantafyllopoulos and Schuller, 2024), coupled with the explosion of available large language models (LLMs), will substantially empower bad actors who intend to use them for vishing. These generative methods can be employed to rapidly increase the amount of vishing attacks by substituting human attackers with automated conversational agents that can perform this act at scale. A recent estimate places fraud at 18% of generative AI misuse (Marchal et al., 2024). While the speech community is scrambling to put countermeasures into place, as seen in the latest anti-spoofing challenges (Liu et al., 2023a) and LLM 'watermarking' efforts (Kirchenbauer et al., 2023), detection performance of synthetic speech is lagging behind, while synthesis fidelity and quality is progressing in leaps and bounds (one need only look at the advances to be presented in the field's conferences and journals in the last few years). While well-intentioned researchers and practitioners strive to adhere to ethical and legal requirements about the use of their research, and try to raise awareness about the potential misuse of speech synthesis, bad actors are not limited by the same considerations. Hence, we can reasonably expect an upcoming "epidemic" of AI-driven vishing attempts, with artificial social engineers who are indistinguishable from humans regarding speech quality and feature similar affective affordances.

As a result, there is a compelling and urgent need to identify vishing attempts through other means. As we discuss below, there are several behaviours at play during a vishing attack. Ranging from deception to persuasion and adapting to the interlocutor, social engineers employ a set of techniques that influence their only means of communication during an attack – voice. At the same time, the victim's voice contains rich information about their current emotional and cognitive state, which indicates their potential susceptibility to an attack. This raises the question of whether linguistic and paralinguistic cues can be utilised to detect a vishing attack as it happens — both from humans and AI agents.

Most importantly, the real-time and often fast-paced nature of vishing attacks, the complicated nature of the phenomenon, and the (by construction) less-than-perfect accuracy of statistical learning algorithms that are typically involved in the detection of speech behaviours, raise a fundamental question of how to deal with those attacks once detected.[2] Specifically, given that all algorithms encapsulate an amount of uncertainty related to their decision and a non-negligible margin of error, the question of *interventions* is non-trivial. It is important to know when, how, and how often to intervene to sufficiently protect users and simultaneously avoid 'warning fatigue' (Distler et al., 2020). This calls for an integrative solution that relies on speech technology for a (non-perfect) detection procedure and intelligent interfaces to communicate that decision to a prospective victim.

Our literature review has revealed an apparent lack of focus on vishing on behalf of the speech community, with very few papers dedicated to this topic and an additional lack of available datasets. While substantial work does exist on the constructs underpinning vishing attacks, there has been very little emphasis on integrating these components and employing them in real-life vishing scenarios. This dearth of available literature has led us to opt for a narrative, rather than a systematic review. Moreover, we have followed an interdisciplinary approach, with the goal to bridge the gap between the community studying vishing from a psychological perspective, and the speech community spearheading the research that can be employed to identify vishing attacks in practical scenarios. Given the lack of a 'solution' to this pressing problem, we finish our work by proposing a roadmap for the development of efficient interventions in. Section 5

In the remainder of this work, we present the key strategies employed by SE attackers in the case of vishing (Section 2) and connect them to the existing research on the social markers they rely on from the perspective of speech technology (Section 3). Following that, we present an overview of how generative methods may soon achieve the capabilities to conduct such vishing attacks under the orchestration of a malicious actor (Section 4). Finally (Section 5), we discuss the challenges that arise from the subversive nature of vishing and lay out a roadmap for how speech technology advances need to be combined with intelligent interfaces to combat this pressing societal challenge.

[1] www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf

[2] To illustrate how pressing this issue might get, we consider the case of SE attacks in general. With 300,000 yearly attacks in the US alone and a 99.5% detection accuracy, we will have 1500 undetected attacks that remain unnoticed. Given the fact that the most sophisticated attacks will be aimed at bigger targets (e.g., banks or other important enterprises), this might still result in a substantial security risk.

2. Background

In the present section, we discuss relevant background for vishing, beginning with the foundational principles underlying social engineering attacks in general. We also outline how a real-time intervention against a detected vishing attack can look like, which forms the bridge with the next section discussing how vishing attacks take place in actual phone conversations.

2.1. Prevalence of vishing

It is hard to estimate the exact extent of vishing. One reliable source of information is the Consumer Sentinel Network operated by the U.S.A. Federal Trade Commission which compiles yearly reports of fraud incidence. Its 2023 report[3] catalogued 2,566,261 fraud reports, of which 1,515,107 reported the contact method. 20% of total fraud attempts were performed via phone call, and 90% of those resulted in a financial loss. The total loss amounted to $850M, with a median $1480 loss per case. The most prevalent type of fraud attempts were imposter scams, followed by online shopping, investment or business related attempts, prizes, internet and telephone services and others. This report gives only some idea of the problem at a global scale and is only based on reported attempts. Crucially, there is no existing technology that can safeguard vishing attacks (as opposed to phishing, which can be filtered by automated methods). The most common 'defense' is exposing caller identity, in the hopes that this will help the prospective victims verify the identity of the caller. Even without considering its effectiveness, this defence is primarily relevant for identity theft attacks. This is what makes vishing a very pressing societal problem with no apparent solution.

2.2. Psychological processes underlying social engineering

Social engineering attacks can be seen as a type of psychological attack that attempts to persuade a victim to complete actions that are beneficial to the attacker (Montañez et al., 2020). These attacks typically follow a common pattern. The attacker collects information about the target, develops relationships with the target, exploits available information to execute the attack and persuade the victim, and exits with no (or little) traces (Junger et al., 2023).

The elaboration likelihood model (Petty et al., 1986; O'Keefe, 2013) was created to explain how people process persuasion stimuli, and might be a promising approach to explain social engineering on a psychological level. In everyday life, people are exposed to many types of persuasive messages (e. g., marketing, other people) and would not be able to carefully scrutinise every message they receive (Petty et al., 1986). The elaboration likelihood model posits that successful persuasion of a person depends on the extent to which they think about ("elaborate") information relevant to the persuasive issue. The model suggests that there are two kinds of persuasion processes that can be engaged. One process involves systematic thinking (*central route*) and the other involves cognitive shortcuts (*peripheral route*). If a person is persuaded using the "central route", they carefully examine the issue and engage critically with the information provided to persuade them. This is typically not the primary route of persuasion engaged in social engineering attacks. On the other hand, the peripheral route refers to a persuasion process when less systematic thinking is involved, and the person being influenced relies on peripheral cues. Instead of carefully considering the issue somebody tries to persuade them of, the person being influenced might use heuristics to guide attitude and belief, for instance how credible the attacker seems to be (O'Keefe, 2013). Many factors influence whether a person engages critically with the information provided by a person who tries to persuade them, including distraction, personal relevance, personal responsibility, and need for cognition (Petty et al., 1986). For example, a vishing target who is currently distracted and who finds the attack pretext to be of little personal relevance might not critically scrutinise the pretext and notice discrepancies that point to a vishing attack. Applying the lens of the elaboration likelihood model to vishing is promising as it provides a framework through which to understand how the attacker manages or fails to persuade the victim.

2.3. Interventions against vishing

A conceptual overview of the interplay between attacker and victim is shown in Fig. 1. As we discuss below, the attacker employs a set of strategies, while the victim exhibits a set of behaviours that indicate their engagement with the attack. The goal of a vishing prevention system should be to intervene during the process of a conversation and notify the prospective victim of an ongoing attack. Importantly, this intervention must be done in real time and *before* the attack succeeds. This can be achieved by an application 'listening in' on all calls on a user's phone and 'pinging' the user whenever it detects a malicious attack. However, the interventions must be kept to a minimum and should be triggered only when a vishing attempt is detected. Moreover, not all attempts are successful; the victim might identify one independently. Thus, an intervention is needed only when an attack is ongoing *and* the victim is 'falling for it'. This requires an analysis of both the caller and the receiver — and their interaction.

As shown in Fig. 1 and discussed in previous work (Jones et al., 2021), an attacker might co-opt a set of persuasion techniques to achieve their goal. Further, throughout the attack, they are engaged in an act of deception. Both those cues can influence their voice. On the receiver end, a successful attack is denoted by increased engagement, while on a conversational level we expect to see an increased interpersonal adaptation between the two interlocutors. All these factors become important for identifying an ongoing and potentially dangerous vishing attack. In the following section, we give an overview of prior research aimed at detecting each of those conversational aspects *in isolation*.

[3] https://www.ftc.gov/system/files/ftc_gov/pdf/CSN-Annual-Data-Book-2023.pdf

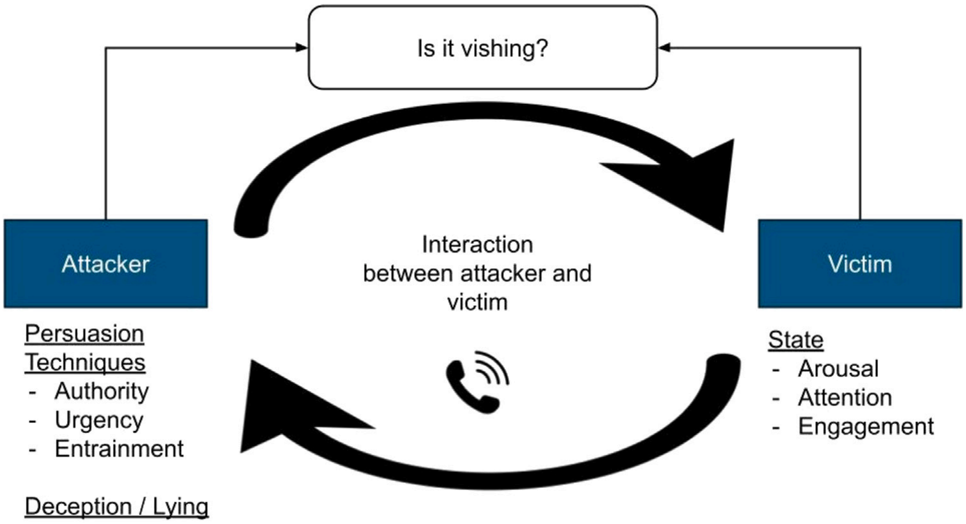

Fig. 1. Conceptualisation of the interaction between attacker and victim, and observable indicators on both the victim and the attacker side.

3. Vishing effects on conversation

This section outlines the effects of a vishing attempt on a conversation. We begin with the attacker, discussing how deception influences their voice, continuing with the impact of persuasion, and, finally, the discrepancy created by trying to maintain the facade. For victims, we mainly focus on engagement, as that is a major telltale sign of the imminent success of an attack. Finally, we discuss the role of entrainment in increasing the likelihood of success.

3.1. Deception

Lies and deception are preeminent in vishing attacks. Automatic lie detection based on vocal cues has been widely studied in the context of criminology and forensics (Pérez-Rosas et al., 2015a), financial misconduct (Hobson et al., 2012), or daily-life occurrences of lying (Scherer et al., 1985). Therefore, it serves as the primary entry point for speech technology research to mitigate vishing. Deception manifests both in the linguistic and the paralinguistic streams of speech – and in the facial expressions and gestures of the subjects, which we ignore here, as this information cannot easily be captured during voice calls.

The above-mentioned cues have been used to classify speech automatically as honest/sincere vs deceptive. However, one significant challenge currently faced by the field is the ecological validity of these findings and their translations to new contexts like vishing. Nahari et al. (2019) highlight the difference that culture or different setups might make on lie detection tools for security applications. In that light, vishing might require a different treatment and a targeted data collection protocol than deception detection in other domains. We discuss important aspects in the following.

3.1.1. Deceptive speech data

A crucial first step in building a speech-based deception detection algorithm is the collection of representative data. The major challenge lies in determining the ground truth. The whole point of deception is to remain hidden; thus, it is often required to ascertain the veracity of a given statement or narrative post-hoc. Another challenge is the scarcity of publicly available datasets, as many studies do not release their data. Table 1 provides an overview of representative datasets, the context in which they took place, and whether they are publicly available. We proceed to analyse their key facets and highlight their differences regarding vishing.

3.1.2. Deception scenarios

The easiest approach is to 'elicit' a lie in a *controlled setting*.

Instructions Elicitation in controlled settings is often done by *outright instructing or indirectly incentivising* study participants to lie. For instance, the data used in Schuller et al. (2016) was collected by having college students obtain a token via theft and then hide that fact when questioned about it. In Hobson et al. (2012), students were incentivised to misreport their performance in a mock exam. Other studies have participants provide both true and false opinions (e. g., Pérez-Rosas et al., 2014) on different topics or tell real and fabricated stories (e. g., Vance et al., 2022).

Table 1
Overview of recent datasets on deception detection in voice, as well as (a) **Scenario:** The underlying context of deception; (b) **Staged:** Whether the dataset was recorded in 'laboratory', staged setup (✓) or recorded in-the-wild (✗); (c) **Availability:** whether data is available (✓) on request (🔒) or not at all (✗); (d) **Duration:** Average duration of each instance; (e) **Modalities:** Whether only text (T), audio (A), or audiovisual (A,V) information is available; (f) **Participants**: Number of participants in the deceptive scenario.

Dataset	Scenario	Staged	Availability	Duration (≈)	Modalities	Participants
POLitical LYing (POLLY) (Bai et al., 2022)	Political statements	✗	🔒	30 sec	A, V	1
DOLOS dataset (Guo et al., 2023b)	Gameshow	?	✓	5 sec	A,V	1
Bag of lies (Gupta et al., 2019)	Description	✓	✓	3.5-42 sec	A,V	1
Columbia-SRI-Colorado (CSC) corpus (Hirschberg et al., 2005)	Performance misreporting	✓	🔒	?	A	1
Daily deceptive dialogue corpus of Mandarin (DDDM) (Huang et al., 2019)	Misreporting	✓	✗	17 min	A	2
Idiap Wolf Corpus (Hung and Chittaranjan, 2010)	Werwolf	✓	✓	6 min	A,V	8-12
Thievery Russian Lie (TRuLie) (Karpova et al., 2021)	Mock crime	✓	🔒	6 sec	A,V	1
Columbia X-Cultural Deception (CXD) Corpus (Levitan et al., 2015)	Misreporting	✓	🔒	45 min	A	2
Miami University Deception Detection Dataset (MU3D) (Lloyd et al., 2019)	Relationships	✓	✓	35 sec	A,V	1
Romanian Deva Criminal Investigation Audio Recordings (RODeCAR) (Mihalache et al., 2019)	Criminal investigations	✗	✓	2 sec	A	1
ReGIM-Lab Lie Detection DataBase (ReliDDB) (Nasri et al., 2016)	Stories	✓	✗	1-2 min	A	1
Multimodal dataset for deception detection (Pérez-Rosas et al., 2014)	Mock crime + best friend + abortion	✓	✗	2-3 min	A,V	1
Real-Life Trial Deception Detection Dataset (Pérez-Rosas et al., 2015a)	Court trials	✗	✓	28 sec	A,V	1
Real-Life Deception Detection Dataset (Pérez-Rosas et al., 2015b)	Street interviews	✗	✗	27 sec	A,V	1
Deceptive Speech Database (DSD) (Schuller et al., 2016)	Theft	✓	🔒	2 min	A	1
Box of Lies (Soldner et al., 2019)	Gameshow	?	✓	6 min	A,V	2
Deception Detection and Physiological Monitoring (DDPM) dataset (Vance et al., 2022)	Experience + travel screening + opinion	✓	✓	11 min	A,V	1
Indonesian Deception Corpus (IDC) (Warnita and Lestari, 2017)	Performance misreporting	✓	✗	3 sec	A,V	1

Games A more natural way to elicit deception is through games, such as the popular "Werewolf" game, where participants must hide their true identity from others Hung and Chittaranjan (2010). Additionally, Soldner et al. (2019) and Guo et al. (2023b) utilised recorded data from online gameshows where participants had the task of exhibiting deceptive behaviour, including giving false descriptions of objects and misleading others about their intentions. In these cases, the ground truth of the data is readily available, resulting in little ambiguity about the labels.

Alternatively, speech may be collected in a more *naturalistic setting*, and its veracity is subsequently verified by a post-hoc investigation (e.g., Pérez-Rosas et al., 2015a, Bai et al., 2022). This is a common scenario in forensic applications, where suspects are first questioned, and their claims are subsequently scrutinised by criminal investigators. In this case, the labelling is done 'after the fact' and is only valid insofar as the associated proof of conviction or acquittal is too.

3.1.3. Spoken content

A major aspect differentiating the datasets used for deception detection is the type of speech data they contain. Some datasets are based on single statements, which may or may not be related to an event whose truthfulness is or is not known. For example, Schuller et al. (2016) have their college students report about an event that is highly pertinent to the theft. In contrast, Pérez-Rosas et al. (2015a) have collected short snippets of real-life trials on which the outcome is known. Although, it is not verified whether they correspond to veracious or dishonest statements.

Other datasets are, in turn, based on interviews. This is typical in law enforcement applications (Nahari et al., 2019) but can also be found in other domains. For example, Hobson et al. (2012) simulated a short, unstructured interview given to firm executives after a public presentation of financial statements (their goal was to identify fraudulent claims made by said executives). In this case, not all of the questions elicited dishonest responses — a more realistic setup given that not all statements in a deceptive narrative must be untrue.

We thus end up with different forms of speech samples: short statements, perhaps given as a response to targeted interview questions, and longer narratives that intertwine falsehood with potentially true facts. Both are relevant for our targeted scenario of vishing — an attacker might first engage in an initial 'pitch', during which they will present a (largely) false narrative and be subsequently placed under some scrutiny by the victim where they have to answer a set of targeted questions. However, as we argue in the next subsection, no existing dataset matches precisely the vishing scenario.

3.1.4. Vishing data

It is readily apparent that most of these datasets do not fit a vishing scenario as a) they contain one participant making a 'public' announcement, e.g., in court (Pérez-Rosas et al., 2015a), as part of a public finance disclosure (Mayew and Venkatachalam, 2012), or within the context of a laboratory experiment (Hobson et al., 2012; Schuller et al., 2016; Lloyd et al., 2019), b) they are staged, except for (Hobson et al., 2012; Pérez-Rosas et al., 2015a), or c) the raw data are no longer available (for Hobson et al., 2012 analysis features were extracted on-the-fly and the audio data subsequently discarded). Moreover, the context in which these datasets were collected does not match vishing. The only exception is Lee and Park (2023). However, the dataset only contains positive instances of vishing without a control group, thus necessitating the need to incorporate other data sources, which jeopardises the generalisability of any findings given the propensity of AI algorithms to exploit subtle domain mismatch as shortcuts. To the best of our knowledge, no recent dataset, publicly available or otherwise, reports classification results for deception detection in a vishing setup or even a closely related scenario — a two-way conversation over the phone, without visual contact. *This makes collecting one a pressing need for future research.*

3.2. Persuasion

In the previous section, we discussed deception as the main facet of SE attacks. However, SE is a complex interpersonal communication phenomenon that depends not solely on lying (Jones et al., 2021). Social engineers rely on various multifaceted persuasion strategies for circumventing the defense mechanisms of their victims. For instance, by investigating a set of real-life vishing calls, Jones et al. (2021) found that authority (95%), social proof (91%), and distraction (90%) were the most frequently used strategies, followed by liking, similarity, and deception (76%), with the top ones falling under the umbrella of *persuasion* techniques. In the following, we chart out the persuasion components that may manifest in an SE attack and are amenable to detection via speech signals.

3.2.1. Strategy: Authority

The attackers often try to project an image of authority, e. g., by claiming to be part of a specific institution or having authorised access to private information from the victim. The main motivation behind this strategy is to elicit the victim's trust by simulating a knowledgeable or experienced interlocutor. The projection of authority through voice cues has been consistently linked to lower pitch (Tigue et al., 2012; Sorokowski et al., 2019). However, this marker is entirely speaker-dependent: a speaker may project power by lowering their pitch, but, to detect that, one needs a frame of reference for this speaker which is generally unavailable.

To our knowledge, no prior work has focused on the automatic classification of authoritative speech (though dominance is a related concept that has received considerable attention Schuller, 2018). There is substantial related literature in detecting leadership attributes (Schuller et al., 2022). However, authority emanating from a social position is a different construct from the one we consider here; there, authority stems from a leader's position (e. g., a politician), whereas for vishing, authority is defined as being in a position of importance for a very specific, niche topic (e.g., an IT representative trying to access a password). The attributes related to leader speech (charismatic, inspiring, etc.), might thus not be related to the type of authoritative speech relevant here (Schuller et al., 2022).

3.2.2. Strategy: Urgency

This strategy aims to overload the victim with information or evoke strong emotions, such as by introducing an element of time pressure with respect to an action that needs to be taken ("This offer is only available for a limited amount of time"). In terms of changes to the voice, it has been shown that people tend to speak louder, at a higher frequency, and with a broader pitch range when speaking urgent words compared to nonurgent or monotone words (Hellier et al., 2002). Additionally, urgency can also be detected on the receiver end, with a higher average fundamental frequency and speech rate having been shown to occur with increased felt urgency (Jang, 2007).

3.3. Challenge: Internal representation discrepancy

So far, we have exclusively discussed verbal cues that may betray a vishing attack from a communicative perspective. However, a social engineer in the process of an attack can be alternatively seen as a speaker in a particular affective state, namely that of lying. While Zuckerman et al. (1981) argued that lying could not be attributed to one definitive behavioural cue, they nevertheless proposed that somebody telling a lie will exhibit increased arousal, specific affective responses that emerge as a corollary of lying (such as anxiety), attempt to control their interlocutor (i. e., authority), and exhibit increased cognitive load in their attempt to maintain the facade. This stems from the fact that the attacker must present a representation of the world to their victim that is incongruent with their internal representation (DePaulo et al., 2003). This set the stage for later work, which delineated deception as a behaviour that induces a cognitive discrepancy on its subject (DePaulo et al., 2003); in plain words, this means that somebody telling a lie must present a representation of the world to their victim that is not congruent with their internal representation. This discrepancy manifests both in the linguistics and paralinguistics; for example, from a linguistic perspective, lies have a more shallow linguistic structure than veracious statements (DePaulo et al., 2003).

Both cognitive load and increased arousal states have been extensively studied regarding their impact on speech. Arousal, in particular, falls under the auspices of the much broader speech emotion recognition (SER) research (Schuller, 2018), with its recognition particularly conducive to audio and paralinguistics. This can be supplemented by recent advances in valence recognition (Wagner et al., 2023), which may further help distinguish between positive and negative high arousal states, with negative being better indicators of anxiety on the part of the attacker, as opposed to positives which might be indicative of an attempt to influence the victim.

Cognitive load and stress have also been shown to impact vocal expressions (Baird et al., 2021). Artificially induced stress manifests as increased pitch variability and voice intensity, with speech samples predictive of cortisol spikes and physiological parameters such as heart rate and respiration rate (Baird et al., 2021). Cognitive load, in turn, has also been shown to increase both absolute pitch and intensity (Huttunen et al., 2011). However, these markers are generally also associated with increased arousal (Schuller, 2018). This makes it hard to identify each of those factors in isolation, especially in a multi-turn conversation where they all interact both within, and across turns. In summary, we expect the interplay between all these factors, namely increased arousal, negative affect, cognitive load, and stress, to manifest in the speaker's voice and thus be detectable using speech-based models. As a result, they can be used as additional indicators of a vishing attack.

3.4. Reaction: Engagement

So far, we have mostly focused on vishing's effects on the attacker's voice. However, when aiming for real-time interventions, monitoring the (prospective) victim's reaction is important. An engaged listener will be more susceptible to getting duped by the attacker. Thus, detecting engagement can indicate that a vishing attack is successful and an intervention is warranted. Engagement can be more easily detected using visual cues, while acoustic detection has proven more challenging (Huang et al., 2016), and thus requires more work to deploy in a real-life setting. In general, we expect an association between engagement and arousal, with the latter serving as a proxy to the former, and being easier to detect (Schuller, 2018). Nevertheless, engagement is still open and – as with all other topics discussed – especially so in a vishing context.

3.5. Conversation: Speaker entrainment

Finally, the interplay between attacker and victim is important to identify potentially successful vishing attacks. Interpersonal adaptation, or *entrainment*, is another strategy social engineers employ. In our context, entrainment is a strategy intentionally pursued by the attacker rather than naturally emerging in interpersonal communication. When appropriately instigated, this "chameleon effect" may increase the likability of the attacker (Chartrand and Bargh, 1999), increasing the chances of a successful attack (Jones et al., 2021). Entrainment can be detected both on a lexical (Brennan and Clark, 1996) and a paralinguistic level (Amiriparian et al., 2019). It relies on quantifying speaker similarity (in terms of distance metrics such as Euclidean distance or other), and tracking its change over time. Entrainment cuts both ways. Thus, we may detect an effort by the attacker to adapt to their victim or a subconscious adaptation from the victim itself. In both cases, alerting the potential victim that interpersonal adaptation is increasing in a voice call where other markers of deception are simultaneously active might serve as a warning that they are susceptible to scamming (cf. mitigation strategies below).

4. GenAI and vishing

Having reviewed the impact of vishing on human speech (for the attacker and the receiver), we next turn to how some of those states and traits on the attacker end can be simulated using generative methods. We will focus here on those aspects that are working in the attacker's favour (authority, imparted urgency, persuasion), assuming that a generative model can be controlled to avoid those telltale signals that unwittingly betray an attack (stress, cognitive load). We begin with a discussion of how LLMs can be used to craft a convincing dialogue in Section 4.1 and continue with how that dialogue can be synthesised with the necessary paralinguistic markers in Section 4.2.

4.1. Vishing and LLMs

LLMs are neural networks trained to predict the probability distribution of the next word in a sequence of words, given the previous words. Mathematically, this can be expressed as $P(w_t|w_1, w_2, \ldots, w_{t-1}; \Theta)$, where Θ is the set of parameters/weights to be learned, and the set of $(w_i)_i$ forms the *prompt*. The words are also named *tokens* and correspond to vector representations. By extension, w_i could even represent audio or other modalities. By training the models on a large corpus of data, the parameters Θ are adjusted to minimise the difference between the predicted probabilities and the actual observed frequencies of token sequences in the training data. Undoubtedly, LLMs have ushered in a new era of AI-driven communication, with applications spanning diverse domains (Kumar et al., 2023; Hou et al., 2023). Such instances include customer service, content creation, and personal assistants. However, the same technology that offers these benefits can also be exploited for malicious purposes. Among the most concerning applications is the use of LLMs in cyber attacks. Leveraging their sophisticated capabilities, cybercriminals can craft highly convincing and adaptive scripts that mimic human interaction with alarming accuracy (Yao et al., 2024; Kumar, 2024). In the case of contextuality and persuasion, LLMs are powerful tools for attackers to manipulate their victims with unprecedented effectiveness. While traditionally, attackers rely on predefined scripts and their own persuasive skills to deceive victims, LLMs and, by extension foundation models (FMs) can amplify the threat by generating dynamic and context-aware responses. We saw that the generation process in LLMs can be conditioned to precede cues, which allow them to tailor the process to the victim's real reactions, leading to highly realistic interaction and authenticity. These advanced models are extended to audio understanding and generation, leading to what is known as large audio models (LAMs) (Triantafyllopoulos et al., 2024). As these models improve in generating natural and coherent speech, the distinction between genuine and fraudulent communication will become increasingly unclear. This underscores the need for suitable intervention strategies.

To illustrate the potential misuse of LLMs in vishing, we consider the following example prompts and the corresponding outputs, which could be easily generated by a state-of-the-art language model.:

Prompt and Output:

```
Create a script for an IT support call informing an employee of a security breach and instructing them to provide
     their login credentials to secure their account.
```

```
Hello, this is Alex from IT support. We've detected unusual activity on your account and need to take immediate
     action to secure it. Could you please provide your login credentials so we can verify your identity and
     implement the necessary security measures? We must address this issue promptly to prevent any potential data
     loss.
```

Prompt and Output:

```
Generate a phone conversation where a bank representative warns a customer about a suspicious transaction and asks
     for verification details.
```

```
Good afternoon, this is Sarah from [Bank Name]'s fraud department. We've noticed a suspicious transaction on your
     account and need to verify some details to ensure your funds are secure. Can you please confirm your account
     number and the last transaction you made? This will help us block any unauthorised access immediately.
```

4.1.1. Prompting LLMs

LLMs, such as GPT-4, are trained on extensive multimodal datasets comprising diverse data from the internet. This training enables them to understand and generate human-like content based on the prompts they receive. However, prompting involves providing them with specific input queries or statements to elicit desired responses, a nuanced and critical task better known as *prompt-engineering*.

Generally, LLMs undergo a multi-stage training process.

1. *Pre-training* involves self-supervised training on a vast corpus of text and audio to understand language patterns, grammar, facts about the world, acquiring even reasoning abilities (Liu et al., 2023c).
2. *Fine-tunning* enables the model to adjust on a narrower dataset with more specific, well-drafted instructions, improving its ability to follow directives and generate relevant responses (Peng et al., 2023).
3. *Prompt Engineering* refers to crafting prompts to guide the LLMs in producing the desired output. Effective prompts are clear and context-rich and can further drive the emerging capabilities of these models (Giray, 2023; Sahoo et al., 2024).

While LLMs are equipped with guardrails to prevent the generation of harmful or illegal content (Banerjee et al., 2024), skilled users can sometimes circumvent these protections through advanced prompting techniques. This manipulation can be achieved in several ways; such instances are: (1) By framing questions or statements in a way that is not overtly illegal or harmful, i.e., instead of asking for illegal instructions directly, such a case would be framing a query in a hypothetical or academic context (Dong et al., 2024). (2) Presenting a detailed scenario where harmful actions are embedded within a broader, seemingly innocent narrative (Yuan et al., 2024). (3) Starting with innocuous questions and gradually introducing more specific and leading prompts (Dong et al., 2024).

The attacks can be roughly classified into *Black-box* and *White-box* attacks. Instances of *Black-box* attacks are *token manipulation*, and *jailbreak prompting* (Zhang et al., 2024b), where the attackers can only access the model through an API. White-box attacks assume that there is complete access to the model (Liu et al., 2024). An instance of this category is *Gradient based attacks* (Kumar, 2024), which gradually introduces sophisticated perturbations in the input with the gradient of the loss until the input is misclassified. These attacks exploit the model's internal structures, akin to the projected gradient descents (PGDs) approach detailed by Geisler et al. (2024), which demonstrates the efficacy and efficiency of continuous relaxation techniques for adversarial prompt generation in large language models.

4.1.2. Persuasion and LLMs

LLMs have shown potential in both detecting fraudulent activities (Wu et al., 2023; Koide et al., 2024) and enhancing their execution through persuasive messaging (Ferrara, 2024; Salvi et al., 2024). Their ability to generate highly convincing and contextually appropriate language opens the door to scaling up misinformation, unwarranted persuasion, manipulation, and outright fraud. Combined with voice cloning technology, which is discussed in the next section, LLMs have raised alarms due to their misuse in impersonating family members or authority figures to deceive victims (Hackenburg and Margetts, 2024). Studies indicate that text generated by LLMs can sometimes be more persuasive than human-written text (Durmus et al., 2024). This suggests that AI can be effectively utilised to persuade people. This, in conjunction with expressive speech synthesiss (ESSs) technology, via adding elements like tone, pitch, and rhythm, makes the afterwards generated speech even more effective. While the pre-training process of LLMs entails intense compute power, efficient fine-tuning methods (such as instruction fine-tuning and RLHF with LoRA (Hu et al., 2021) and its variants) allow the developers to tailor messages to individual demographics or opinions. This is critical for persuasion as authors in Salvi et al. (2024) showed that personalised messages crafted by LLMs surpass non-personalised ones across various domains and psychological profiles regarding influence. This is a consequence of the fact that LLMs can leverage microtargeting to adapt their arguments more effectively than humans for their advantage in personalised settings.

4.2. Vishing and voice synthesis

Following the creation of an 'attack recipe' and its implementation using LLMs as discussed above, the final step for a fully-autonomous vishing agent is using TTS technology to synthesise speech. Such technology already exists for creating high-fidelity speech with very low latency (Triantafyllopoulos et al., 2023). In the present, we focus on how this speech can be made to resemble that of a particular speaker (i. e., voice cloning) and how it can be manipulated to showcase the states and traits that have been proven successful during vishing attacks (and which we covered in Section 3 for human attackers).

4.2.1. Voice cloning

Voice cloning is the process of creating a digital replica of a human voice by analysing and reproducing its unique sound patterns (Triantafyllopoulos et al., 2023). This technology leverages advanced algorithms to capture the distinctive characteristics of a specific voice, resulting in a digital voice nearly indistinguishable from the original. Voice cloning has significant applications in various fields, including personalised virtual assistants, automated customer service, and entertainment.

A related but distinct technology is voice conversion, which involves transforming one person's voice to sound like another's without necessarily creating a digital replica. Voice conversion focuses on altering specific aspects of the voice to match another speaker while retaining the linguistic content.

Voice cloning methods can be broadly classified into two categories:

1. Auto-regressive (AR) Approaches: Techniques such as VALL-E (Wang et al., 2023) extract speaker embeddings or acoustic tokens from the reference audio. These embeddings are then used as conditions for the audio encoder, which generates audio embeddings sequentially.
2. Non-autoregressive (NAR) Approaches: Methods like YourTTS (Casanova et al., 2022) and Voicebox (Le et al., 2024) also use speaker embeddings from reference audio but generate the entire output sequence in a single forward pass, significantly improving inference speed.

Despite the advancements in voice cloning technology, several challenges remain, including ensuring ethical use, improving robustness against misuse, and addressing technical limitations in accurately capturing and reproducing diverse voice characteristics.

4.2.2. Expressive speech synthesis

In recent years, significant improvements have been achieved in the intelligibility and naturalness of synthesised speech on neutral TTS (Tan, 2023). Despite these advances, the expressiveness of the synthesised speech remains a challenge (Triantafyllopoulos et al., 2023). The goal of expressive speech synthesis (ESS) is to generate human-like and natural speech, which is determined by multiple factors such as prosody, emotion, timbre, and content (Chan et al., 2022; Tang et al., 2024). ESS inherently involves a TTS component, as producing expressive speech also requires generating intelligible speech that conveys both linguistic and stylistic meaning (Triantafyllopoulos et al., 2023; Triantafyllopoulos and Schuller, 2024).

One of the key challenges in ESS is addressing the one-to-many mapping problem, where the same input condition can correspond to multiple variations in speech, such as pitch, duration, emotional style, and sound level (Guo et al., 2023a; Yang et al., 2024). Conventional TTS systems commonly leverage pre-defined speaker-ID, style embedding, or emotion embedding extracted from reference speech as conditions to model speech variability. However, these methods are user-unfriendly and lack generalisability (Yang et al., 2022; Kaur and Singh, 2023).

Inspired by the significant improvements and success in text and image generation guided by text descriptions as prompts (Ramesh et al., 2022; Wang et al., 2024), prompt-guided TTS (Guo et al., 2023a; Liu et al., 2023d) and diffusion-model-based TTS (Jing et al., 2023; Popov et al., 2021) have emerged to provide more dynamic and contextually appropriate variations in synthesised speech. Generally, prompt-guided TTS systems are trained with a text prompt dataset, consisting of speech and its corresponding text prompt. Expressive speech is generated by conditioning the model on both the text content to be synthesised and the text prompt describing the variability or style of the voice. Since the text prompt is in natural language, for example: "A lady whispers to her friend slowly", this approach results in significant improvements in audio naturalness and expressiveness (Guo et al., 2023a; Liu et al., 2023d).

4.3. Deepfake detection

Finally, we focus on how AI-generated speech and text can be detected automatically. This pertains to detecting such content irrespective of whether it constitutes a vishing attack or not. Such algorithms have drawn increasing attention in contemporary research due to the proliferation of AI-generated content. Therefore, we discuss them in the context of AI-based vishing. We note that this is not sufficient to denote a malicious attack. For example, AI 'robocalls' can be used for legitimate marketing purposes. Detecting that an artificial agent is making a specific call can only be a first step in deciding whether it constitutes an AI-driven vishing attack.

We begin with a discussion of how content generated by specific models can be watermarked by their creators (to prevent the malicious use of those models) and continue with an outline of how AI-generated text and speech can be detected 'in-the-wild'.

4.3.1. Watermarking models

Watermarking LLMs is a strategy for mitigating the potential harms of these models, such as their use in misinformation and fraudulent activities (Kirchenbauer et al., 2023; Pang et al., 2024). This technique involves embedding hidden patterns into generated text, making it algorithmically detectable without human recognition. The watermark is created by promoting a randomised set of 'special' tokens during text generation, which can be later identified using statistical tests. Early approaches to watermarking natural text, such as those proposed by Atallah et al. (2002), focused on creating watermarked text that retained similar meaning while being detectable only by those with knowledge of the watermarking key.

Modern neural language models have advanced these techniques, enabling more sophisticated and less intrusive watermarking methods. For example, Fang et al. (2017) introduced a neural steganography approach that partitions the vocabulary of a language model to encode messages without significantly degrading text quality. Despite these advancements, challenges remain, such as watermarking low-entropy sequences where human and machine-generated texts are nearly indistinguishable. Moreover, ongoing efforts are to enhance watermark robustness against adversarial attacks, where attackers attempt to remove the watermark by modifying text. Additionally, the development of private watermarking methods, which use cryptographic techniques to keep watermarking keys secret, further complicates the removal process for attackers. These methods aim to balance the need for effective watermarking with minimal impact on the quality of the generated text, making it a practical tool for combating the misuse of LLMs. As the field progresses, further research is required to address the open questions and optimise watermarking techniques for real-world applications.

Similarly, audio watermarking is the process of embedding specific information within an audio signal, imperceptible to humans but strongly detectable by algorithms. This technology can date back to the 1990s, originally developed for copyright protection (Cox et al., 1997) and authentication (Boney et al., 1996).

Traditional methods for audio watermarking relied on embedding watermarks in either the time or frequency domains by leveraging domain-specific features (Cvejic and Seppanen, 2004; Yeo and Kim, 2003; Bender et al., 1996). However, these approaches depended heavily on expert knowledge and empirical rules, resulting in limited encoding capacity and vulnerability to attacks (Dutta et al., 2020).

The advent of deep neural networks (DNNs) has revolutionised watermarking techniques, showcasing significant potential in both encoding and detection (Liu et al., 2023b; Pavlović et al., 2022; Chen et al., 2023; San Roman et al., 2024). Generally, deep-learning methods employ an *Encoder-Attack Simulator-Decoder* architecture, which enhances robustness against predefined attacks and reduces the complexity of encoding strategies through automatic learning. There are two primary types of audio watermarking: non-localised and localised. Non-localised methods embed the watermark throughout the entire audio signal (Liu et al., 2023b; Pavlović et al., 2022), while localised methods divide the audio into smaller segments, embedding a complete watermark within each segment (Chen et al., 2023; San Roman et al., 2024). The key difference between these methods is their ability to identify small segments of AI-generated speech within longer audio clips.

4.3.2. Detecting AI-generated text

The rapid expansion of AI-created content has made deepfake detection a critical priority, and it is already forming a significant downside of the AI explosion and a multidimensional challenge for the future. Traditional methods involve machine learning algorithms that analyse linguistic patterns and inconsistencies in the text. For instance, models like GPT-4 have been evaluated for their ability to detect AI-generated text based on stylistic and contextual anomalies.

Research highlights several key approaches for detecting text deep fakes:

Linguistic Analysis Techniques such as frequency analysis and syntactic parsing identify irregularities in sentence structure and word usage typical of AI-generated content (Wang, 2024).

Residual Analysis Methods of this category aim to detect deepfakes by analysing residuals from biological signals and other subtle inconsistencies (Ciftci et al., 2021).

Multimodal Integration Combining text analysis with other modalities, such as image or video data, can enhance detection accuracy. For example, LLMs can evaluate the coherence between text and associated media, identifying mismatches indicative of deepfake content (Zhang et al., 2024a; Yu et al., 2023).

Contextual Prompts Effective use of contextual prompts in LLMs improves their ability to detect text anomalies. Advanced prompting techniques, such as chain-of-thought or few-shot prompting, guide models to focus on specific inconsistencies within the text (Su et al., 2024).

Despite these advancements, challenges remain. LLMs can generate highly coherent text, making detection increasingly difficult. Furthermore, the reliance on semantic cues necessitates continuous refinement of detection models to keep pace with evolving AI capabilities.

4.3.3. Detecting AI-generated speech

As the threat of audio deep fakes continues to emerge, a series of challenges, such as the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof) (Liu et al., 2023a) and the Audio Deepfake Detection (ADD) Challenge (Yi et al., 2023; Jiangyan et al., 2022), have been established to foster research in this critical area. Since its inception in 2015, the ASVspoof initiative has aimed to promote the development of countermeasures (CM) to address the threat of spoofing attacks on ASV systems. The data for ASVspoof was collected from a broad range of state-of-the-art voice conversion (VC) and TTS systems, as well as various recording and replay devices. As audio deepfakes have become a focal point, the DeepFake Detection sub-challenge was introduced in 2021 (Liu et al., 2023a). The ADD Challenge aims to simulate real-life scenarios, enhancing the robustness of detection systems. The improvements in the ADD Challenge are twofold: (1) incorporating diverse background noises and disturbances in fake audios, and (2) introducing partially spoofed audio data by embedding several small fake clips within real speech audio, rather than using fully synthetic audio.

Concurrently, researchers have been developing methods to detect audio deep fakes and distinguish them from human speech. Detection methods generally fall into three main categories: (1) Low-level approaches that identify artifacts introduced by generative AI (Chakroborty et al., 2008; Hasan et al., 2013; Wu et al., 2013; Wang et al., 2015), such as magnitude-based and phase-based spectral coefficients, which have long been strong baselines for audio deepfake detection; (2) High-level approaches that analyse more complex features like the semantic content of the speech (Yu et al., 2017; Zeinali et al., 2019; Wang and Yamagishi, 2021; Li et al., 2023; Ma et al., 2023); and (3) Approaches that utilise acoustic and linguistic features, analysing the physical properties of the speech signal (Anthony et al., 2016; Dhamyal et al., 2021; Zhang et al., 2023; Fan et al., 2023), such as silence patterns, sub-band frequencies, and prosody and semantic features. In recent years, deep embedding representations, particularly those derived from pre-trained self-supervised learning (SSL) based models, have gained prominence. These embeddings capture a wide range of speech characteristics, enabling more sophisticated analysis and improving the efficacy of detection methods.

5. Research challenges and roadmap

As discussed, vishing has received little attention in speech technology research. In the present section, we argue that the creation of speech-informed protective interventions entails more than training a model to ‘simply’ detect ‘deception’ using existing data, as, due to the need for ecological validity, this model would not easily transfer to the vishing scenario. Moreover, we argue for a need to treat vishing holistically, covering the entire gamut of affective behaviours at play for both the victim and the attacker. This is especially needed for AI-driven attacks, where the ever-improving expressive capabilities of generative AI models may create a cat-and-mouse game, making the detection of an attack increasingly harder.

Vishing detection: The literature distills vishing attacks as a complex set of behaviours encompassing various cues. The core challenge remains that single cues are usually insufficient to prove an attempt; instead, they must be accounted for collectively and integratively. An increased arousal, a projection of authority, and an increased cognitive load may simultaneously describe a speaker not presently engaging in a social engineering attack (DePaulo et al., 2003). For example, business ‘pitch’ calls entail characteristics of social engineering.

Crucially, contemporary speech technology research rarely adopts such an integrative approach. While numerous attempts have been made to detect arousal, cognitive load, stress, or any cues that could signify a vishing attack, these tasks are consistently tackled only in isolation. This is despite the fact that these affective states share a subset of cues that would make it harder to recognise them in parallel. Furthermore, vishing is defined through the juxtaposition of behaviour and *intent* – as discussed, a business call might still be just a business call. This raises a poignant question: *Can speech technology, in its present status, tackle the rising challenge of fraudulent social engineering attempts?*

We believe predictive models alone cannot handle this task independently due to the ingrained epistemic uncertainty surrounding vishing attacks. Rather, we envision a collaborative, human-centred approach that uses these predictions to alert victims and make them cognisant of the particular persuasion or deception markers that might be at play, thus intervening – in real-time – to give a 'reality-check' of the caller's request.

Intervention strategies: Designing effective interventions is particularly challenging, especially in the context of real-time conversations where users need to make informed decisions about potential risks. We envision an application that monitors all calls on a user's phone and actively works to prevent vishing attacks (following their permission, see below). If the system detects malicious activity with high certainty, it could automatically terminate the call. However, it is essential to inform the user of the reason for this action, providing a clear explanation of why the call was flagged as suspicious. Alternatively, a more user-centred approach might involve alerting the user during the call and offering actionable strategies to safely end the conversation themselves, thereby fostering user understanding and engagement in the process.

Repeated occurrences of falsely displayed warnings can lead to habituation and warning fatigue (Distler et al., 2020), meaning that technology users start ignoring warnings since they perceive them as likely incorrect. Warning literature often focuses on compliance, with users taking the recommended (usually "safer") path of action, but it is also important to evaluate whether interventions are understandable and how they influence user experience, especially in the presence of false positives (Distler et al., 2020).

Moreover, vishing typically occurs in working environments, where the victim might already be overloaded with work-related tasks and thus have limited attention resources to detect the attack. This means that even when a technology user knows, in principle, about the presence and indicators of certain attacks, contextual factors can lead them to engage in unsafe behaviour (Distler, 2023). It is thus an open question how to best communicate with the user about suspicious spoken communication, taking into consideration their prior knowledge, current state (e.g., emotions, attention, motivation) and other context. While no concrete interventions have yet been proposed for vishing, we hypothesise that drawing inspiration from measures already tested against vishing or other SE attacks should form a starting point for further research.

Classical real-time intervention strategies to counter phishing attempts focus on providing users with in-the-moment support during interactions to improve decision-making and mitigate risks. These include interactive warnings, which interrupt users during potentially unsafe actions, such as clicking a suspicious link or providing sensitive information. Studies have shown that forced-attention mechanisms, such as delayed link activation or interactive prompts (e. g., "Are you sure you trust this email?"), significantly reduce phishing susceptibility compared to passive warnings like static toolbars (Egelman et al., 2008). Highlighting critical elements, such as the domain name or sender's address in emails, is another widely used approach to help users quickly identify suspicious attributes (Volkamer et al., 2016). Dynamic risk indicators, such as traffic-light-style visual cues, provide real-time feedback by assigning risk levels to email or website interactions (Franz et al., 2021). Another promising avenue is *digital nudges*, which gently guide users' decisions by leveraging behavioural heuristics in real-time. Examples include social salience nudges, such as showing how peers reacted to similar emails, or fear-based nudges, which highlight potential consequences of interacting with suspicious content (Nicholson et al., 2019). Recent real-time interventions embed user-centric security features directly into email interfaces to support decision-making. Zheng and Becker (2023) evaluated three tools: a "check" button to verify sender details and past correspondence, a "collegiate phishing report" nudge and a "suspicion score" nudge. Users found the "check" button useful for affirming legitimacy, while the "suspicion score" nudge was most effective in raising vigilance without disrupting workflows.

It is not yet clear how the previous, vision-based interventions would translate to phone calls, but the same underlying principles apply when considering vishing. Integrating real-time interventions into vishing prevention systems can leverage both in-call mechanisms and additional personal devices to provide comprehensive support for users. *In-call interventions* could focus on delivering unobtrusive, context-aware feedback during ongoing conversations. Subtle audio cues, such as beeps or low-volume tones, can alert users to suspicious patterns, such as urgency or manipulative speech, without disrupting the flow of dialogue. Whispered prompts or binaural audio can provide personalised guidance directly to the user, suggesting actions like verifying the caller's identity or refraining from sharing sensitive information. During natural pauses in the conversation, context-aware systems can deliver actionable advice, such as recommending clarifying questions or prompting reflection on the caller's credibility. For high-risk scenarios, the system might interrupt the call with a clear warning or provide an option to terminate it, empowering the user to take control of the situation.

Expanding beyond the phone call, *interventions involving other personal devices* offer additional avenues for real-time support. Smartwatches or smartphones can display visual cues such as a "risk indicator" meter or provide concise explanations of detected threats, such as "Caller exhibits manipulative speech patterns. Verify identity". These devices can also offer actionable advice, like suggesting specific steps to validate the caller's claims or quick access to emergency contacts. Features like a "suspicion score" or past correspondence checks, inspired by recent email security tools could be adapted for call scenarios. Users could access these tools to assess the legitimacy of the caller based on previous interactions or flagged risk indicators.

6. Ethical considerations

In this last section, we lay out some key ethical considerations regarding the development and deployment of vishing detection models and suitable interfaces. These pertain to an infringement on the user's privacy and the prospect of "dual use" for the developed models.

6.1. Privacy

The proposed approach of leveraging information from voice calls to identify and protect from vishing attacks necessitates careful consideration of the implications on privacy. A trade-off between privacy and security is integral to many security mechanisms, especially for biometric authentication mechanisms, where added security is traded for the provider to have access to the user's biometric samples. In the following subsections, we provide a short discussion of three crucial aspects of privacy. While the trade-off between privacy and security is subject to the personal preferences of each individual, we envision *opt-in* protective mechanisms, with users explicitly agreeing to the use of phone call monitoring systems in return for increased safety. Further, we may hypothesise a tiered protection ecosystem, with more holistic – and thus intrusive – monitoring systems activated once coarser models detect an ongoing vishing attack. For example, assuming that users are more mindful of a system that monitors the semantics of conversation, rather than their acoustic tone, we can envision a system which predicts detection attempts based solely on paralinguistic information, with a successful detection triggering the activation of a linguistic module that complements the first module.

6.1.1. Data protection

Technical approaches such as differential privacy (Dalenius, 1977) can protect user data required for training. At the same time, such approaches are difficult for end-users to comprehend. To address this, alternative approaches include providing users explicit control over when and which data is being collected. For example, Buschek et al. (2018) showed how users' privacy concerns could be addressed in the context of keystroke dynamics (that is, authentication approaches analysing users' typing behaviour). Users could explicitly deactivate data collection through a switch integrated with the keyboard. In addition, sensitive information (such as entering passwords) was excluded during keystroke assessment. Another approach is to sample data so that the original conversation cannot be easily reconstructed from the data (e. g., random selection of n-grams). Such approaches are potentially useful in the case of vishing — with only parts of the monitored conversations stored for future auditing. The storing of such data would be solely in the control of the user. For example, following the automatic detection of a vishing attack, a monitoring system could prompt the user to store the parts of the conversation that were identified as being relevant for the attack, which could be used as proof when contacting the authorities.

6.1.2. User acceptance

Closely related is how acceptance rates for such an approach can be increased. Prior work showed that carefully explaining to users the benefits of particular security mechanisms generally increases their motivation (Adams and Sasse, 1999). The apparent challenge is to convey information important for users' decisions comprehensibly, including which information is collected, how it is transmitted, where it is stored, and who has access to it. This calls for an integrative, human-in-the-loop approach to the (co-)creation process of models; concretely, for speech practitioners, it means involving potential users in the creation process *before* they reach a decision on which components of a conversation they are allowed to monitor (for instance, users may be more loathe to allow an algorithm processing their own end of a conversation, thus precluding this part as input for a detection model).

6.1.3. User interfaces for consent and permission control

Privacy user interfaces are required to obtain consent from users and provide them control over how their data is being used. Traditional permission control systems, as known from Android, witnessed a transition from requesting permission before installing an application towards contextual approaches (Prange et al., 2024). For example, current smartphone apps prompt users for consent at the very moment they require access to data (such as location). There is a need for research on adopting such approaches to voice-based user interfaces with the goal of allowing users to make informed decisions while at the same time minimising any negative influence on the objective of protecting users from undesirable actions.

6.2. Dual use

Finally, developers interested in creating vishing detection modules must be aware of their potential for **dual use**. From a technical perspective, impostors might leverage protective mechanisms to refine their attacks to a point where their detection becomes difficult or impossible (i. e., to use the available tools in adversarial fashion in order to improve their attacks). This could become a serious threat with the ability to automate voice-based attacks using deep fakes.

From a user perspective, attackers might similarly exploit behavioural patterns. For example, protective mechanisms might create a false sense of security among users, i. e., as users develop an over-reliance and blindly trust voice content not flagged as malicious. Strategies of attackers might be to overwhelm the system with false requests, ultimately causing warning fatigue among users.

Ultimately, this potential for dual use calls for a more interdisciplinary approach to vishing detection. As mentioned above, focusing only on aspects of detection important are not in themselves adequate to tackle this emerging threat. Rather, algorithm developers need to interact closely with interface designers to create effective and robust intervention mechanisms for the continuous protection of users from malicious intent.

6.3. Legal ramifications

As discussed, we envision solely 'opt-in' systems which should address all legal and privacy considerations as it pertains to the rights of the user of a vishing protection software. However, given that any such proposed system will be monitoring the calls of that individual with other interlocutors, the majority of which will not be engaging in malicious acts, their individual rights to privacy come into play as well. Recent regulation, such as the EU AI Act (The European Parliament, 2023), prescribe a 'right of notice' to all users that interact with an automatic monitoring systems. It is not clear how that information *could* be technically communicated to the other end of a voice call (as this might require a special protocol which transmits this information over standard telephone lines), and even if it *should* be communicated in advance, given that it would alert potential attackers to the employment of such software. It is beyond the scope of this article to address all these considerations. Nevertheless, we note that, as the community gets increasingly involved in the development of this protective net of technologies against vishing, they should integrate different legal and societal stakeholders who should ultimately determine the legal framework that guides the use of this technology.

7. Conclusion

We presented an overview of different factors that are at play during a vishing attack, as well as considerations with respect to the ecological validity of available data. Overall, any application for voice-informed interventions will lie at the intersection of human-machine collaboration. An automatic vishing detection system can only detect cues that indicate an attack. Interpreting and acting on those cues is ultimately the responsibility of the victim. This calls for an integrative approach, using principles from speech technology and human-centred design, to timely, and usefully, intervene in the process of a two-way conversation. We consider this an exciting – and urgently needed – area for future work in speech technology research.

CRediT authorship contribution statement

Andreas Triantafyllopoulos: Writing – review & editing, Writing – original draft, Conceptualization. **Anika A. Spiesberger:** Writing – original draft. **Iosif Tsangko:** Writing – original draft. **Xin Jing:** Writing – original draft. **Verena Distler:** Writing – original draft. **Felix Dietz:** Writing – original draft. **Florian Alt:** Writing – review & editing, Writing – original draft, Project administration, Funding acquisition. **Björn W. Schuller:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge funding from the EU Horizon grant agreement No. 101060660 (SHIFT) and the project Voice of Wisdom and funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

Data availability

No data was used for the research described in the article.

References

Adams, A., Sasse, M.A., 1999. Users are not the enemy. Commun. ACM 42 (12), 40–46. http://dx.doi.org/10.1145/322796.322806, URL: https://doi.org/10.1145/322796.322806.

Amiriparian, S., Han, J., Schmitt, M., Baird, A., Mallol-Ragolta, A., Milling, M., Gerczuk, M., Schuller, B., 2019. Synchronization in interpersonal speech. Front. Robot. AI 6, 116.

Anthony, C., Nelson, A.J., Tripsas, M., 2016. "Who are you?... I really wanna know": Product meaning and competitive positioning in the nascent synthesizer industry. Strat. Sci. 1 (3), 163–183.

Atallah, M.J., Raskin, V., Hempelmann, C.F., Karahan, M., Sion, R., Topkara, U., Triezenberg, K.E., 2002. Natural language watermarking and tamperproofing. In: International Workshop on Information Hiding. Springer, pp. 196–212.

Bai, C., Bolonkin, M., Regunath, V., Subrahmanian, V., 2022. POLLY: A multimodal cross-cultural context-sensitive framework to predict political lying from videos. In: Proceedings of the 2022 International Conference on Multimodal Interaction. ICMI '22, Association for Computing Machinery, New York, NY, USA, pp. 520–530.

Baird, A., Triantafyllopoulos, A., Zänkert, S., Ottl, S., Christ, L., Stappen, L., et al., 2021. An evaluation of speech-based recognition of emotional and physiological markers of stress. Front. Comput. Sci. 3, 750284.

Banerjee, S., Layek, S., Hazra, R., Mukherjee, A., 2024. How (un) ethical are instruction-centric responses of LLMs? Unveiling the vulnerabilities of safety guardrails to harmful queries. ArXiv preprint arXiv:2402.15302.

Bender, W., Gruhl, D., Morimoto, N., Lu, A., 1996. Techniques for data hiding. IBM Syst. J. 35 (3.4), 313–336.

Boney, L., Tewfik, A.H., Hamdy, K.N., 1996. Digital watermarks for audio signals. In: Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems. IEEE, pp. 473–480.

Brennan, S.E., Clark, H.H., 1996. Conceptual pacts and lexical choice in conversation. J. Exp. Psychol. [Learn. Mem. Cogn.] 22 (6), 1482.

Buschek, D., Bisinger, B., Alt, F., 2018. ResearchIME: A mobile keyboard application for studying free typing behaviour in the wild. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, Association for Computing Machinery, New York, NY, USA, pp. 1–14. http://dx.doi.org/10.1145/3173574.3173829, URL: https://doi.org/10.1145/3173574.3173829.

Casanova, E., Weber, J., Shulby, C.D., Junior, A.C., Gölge, E., Ponti, M.A., 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In: International Conference on Machine Learning. PMLR, pp. 2709–2720.

Chakroborty, S., Roy, A., Saha, G., 2008. Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. Int. J. Electron. Commun. Eng. 2 (11), 2554–2561.

Chan, C.H., Qian, K., Zhang, Y., Hasegawa-Johnson, M., 2022. Speechsplit2: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6332–6336.

Chartrand, T.L., Bargh, J.A., 1999. The chameleon effect: The perception–behavior link and social interaction. J. Pers. Soc. Psychol. 76 (6), 893.

Chen, G., Wu, Y., Liu, S., Liu, T., Du, X., Wei, F., 2023. Wavmark: Watermarking for audio generation. ArXiv preprint arXiv:2308.12770.

Ciftci, U.A., Demir, I., Yin, L., 2021. Detection of synthetic portrait videos using biological signals. ArXiv.

Cox, I.J., Kilian, J., Leighton, F.T., Shamoon, T., 1997. Secure spread spectrum watermarking for multimedia. IEEE Trans. Image Process. 6 (12), 1673–1687.

Cvejic, N., Seppanen, T., 2004. Increasing robustness of LSB audio steganography using a novel embedding method. In: International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. vol. 2, IEEE, pp. 533–537.

Dalenius, T., 1977. Towards a Methodology for Statistical Disclosure Control. Statistics Sweden.

DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., Cooper, H., 2003. Cues to deception. Psychol. Bull. 129 (1), 74.

Dhamyal, H., Ali, A., Qazi, I.A., Raza, A.A., 2021. Fake audio detection in resource-constrained settings using microfeatures. In: Interspeech. pp. 4149–4153.

Distler, V., 2023. The influence of context on response to spear-phishing attacks: an in-situ deception study. In: Proc. Conference on Human Factors in Computing Systems. ACM, Hamburg Germany, pp. 1–18.

Distler, V., Lenzini, G., Lallemand, C., Koenig, V., 2020. The framework of security-enhancing friction: How UX can help users behave more securely. In: New Security Paradigms Workshop 2020. ACM, Online USA, pp. 45–58.

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., et al., 2024. Safeguarding large language models: A survey. ArXiv preprint arXiv:2406.02622.

Durmus, E., Lovitt, L., Tamkin, A., Ritchie, S., Clark, J., Ganguli, D., 2024. Measuring the Persuasiveness of Language Models. Anthropic.

Dutta, H., Das, R.K., Nandi, S., Prasanna, S.M., 2020. An overview of digital audio steganography. IETE Tech. Rev. 37 (6), 632–650.

Egelman, S., Cranor, L.F., Hong, J., 2008. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '08, Association for Computing Machinery, New York, NY, USA, pp. 1065–1074. http://dx.doi.org/10.1145/1357054.1357219, URL: https://doi.org/10.1145/1357054.1357219.

Fan, C., Xue, J., Dong, S., Ding, M., Yi, J., Li, J., Lv, Z., 2023. Subband fusion of complex spectrogram for fake speech detection. Speech Commun. 155.

Fang, T., Jaggi, M., Argyraki, K., 2017. Generating steganographic text with LSTMs. ArXiv preprint arXiv:1705.10742.

Ferrara, E., 2024. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. J. Comput. Soc. Sci. 1–21.

Franz, A., Zimmermann, V., Albrecht, G., Hartwig, K., Reuter, C., Benlian, A., Vogt, J., 2021. {Sok}: Still plenty of phish in the sea—A taxonomy of {user-oriented} phishing interventions and avenues for future research. In: Seventeenth Symposium on Usable Privacy and Security. SOUPS 2021, pp. 339–358.

Geisler, S., Wollschläger, T., Abdalla, M., Gasteiger, J., Günnemann, S., 2024. Attacking large language models with projected gradient descent. ArXiv preprint arXiv:2402.09154.

Giray, L., 2023. Prompt engineering with ChatGPT: a guide for academic writers. Ann. Biomed. Eng. 51 (12), 2629–2633.

Guo, Z., Leng, Y., Wu, Y., Zhao, S., Tan, X., 2023a. Prompttts: Controllable text-to-speech with text descriptions. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.

Guo, X., Selvaraj, N.M., Yu, Z., Kong, A.W.-K., Shen, B., Kot, A., 2023b. Audio-visual deception detection: DOLOS dataset and parameter-efficient crossmodal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 22135–22145.

Gupta, V., Agarwal, M., Arora, M., Chakraborty, T., Singh, R., Vatsa, M., 2019. Bag-of-lies: A multimodal dataset for deception detection. In: Proc. CVPR.

Hackenburg, K., Margetts, H., 2024. Evaluating the persuasive influence of political microtargeting with large language models. Proc. Natl. Acad. Sci. 121 (24), e2403116121.

Hasan, T., Sadjadi, S.O., Liu, G., Shokouhi, N., Bořil, H., Hansen, J.H., 2013. CRSS systems for 2012 NIST speaker recognition evaluation. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6783–6787.

Hellier, E., Edworthy, J., Weedon, B., Walters, K., Adams, A., 2002. The perceived urgency of speech warnings: Semantics versus acoustics. Hum. Factors 44 (1), 1–17.

Hirschberg, J., Benus, S., Brenier, J.M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B.L., Shriberg, E., Stolcke, A., 2005. Distinguishing deceptive from non-deceptive speech. In: Proc. INTERSPEECH. pp. 1833–1836.

Hobson, J.L., Mayew, W.J., Venkatachalam, M., 2012. Analyzing speech to detect financial misreporting. J. Account. Res. 50 (2), 349–392.

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., Wang, H., 2023. Large language models for software engineering: A systematic literature review. ArXiv preprint arXiv:2308.10620.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: Low-rank adaptation of large language models. ArXiv preprint arXiv:2106.09685.

Huang, C.-H., Chou, H.-C., Wu, Y.-T., Lee, C.-C., Liu, Y.-W., 2019. Acoustic Indicators of Deception in Mandarin Daily Conversations Recorded from an Interactive Game. In: Proc. INTERSPEECH. pp. 1731–1735.

Huang, Y., Gilmartin, E., Campbell, N., 2016. Conversational engagement recognition using auditory and visual cues. In: Proc. INTERSPEECH. pp. 590–594.

Hung, H., Chittaranjan, G., 2010. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In: Proceedings of the 18th ACM International Conference on Multimedia. MM '10, Association for Computing Machinery, New York, NY, USA, pp. 879–882.

Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., Leino, T., 2011. Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. Appl. Ergon. 42 (2), 348–357.

Jang, P.-S., 2007. Designing acoustic and non-acoustic parameters of synthesized speech warnings to control perceived urgency. Int. J. Ind. Ergon. 37 (3), 213–223.

Jiangyan, Y., Ruibo, F., Jianhua, T., Shuai, N., Haoxin, M., Chenglong, W., Tao, W., Zhengkun, T., Xiaohui, Z., Ye, B., Cunhang, F., Shan, L., Shiming, W., Shuai, Z., Xinrui, Y., Le, X., Zhengqi, W., Haizhou, L., Zheng, L., Bin, L., 2022. Add 2022: the first audio deepfake detection challenge. ArXiv preprint arXiv:2202.08433.

Jing, X., Chang, Y., Yang, Z., Xie, J., Triantafyllopoulos, A., Schuller, B.W., 2023. U-DiT TTS: U-diffusion vision transformer for text-to-speech. In: Speech Communication; 15th ITG Conference. VDE, pp. 56–60.

Jones, K.S., Armstrong, M.E., Tornblad, M.K., Siami Namin, A., 2021. How social engineers use persuasion principles during vishing attacks. Inf. Comput. Secur. 29 (2), 314–331.

Junger, M., Koning, L., Hartel, P., Veldkamp, B., 2023. In their own words: deception detection by victims and near victims of fraud. Front. Psychol. 14, 1135369. http://dx.doi.org/10.3389/fpsyg.2023.1135369, URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1135369/full.

Karpova, V., Popenova, P., Glebko, N., Lyashenko, V., Perepelkina, O., 2021. "Was it you who stole 500 rubles?" - the multimodal deception detection. In: Companion Publication of the 2020 International Conference on Multimodal Interaction. In: ICMI '20 Companion, Association for Computing Machinery, New York, NY, USA, pp. 112–119.

Kaur, N., Singh, P., 2023. Conventional and contemporary approaches used in text to speech synthesis: A review. Artif. Intell. Rev. 56 (7), 5837–5880.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T., 2023. A watermark for large language models. In: International Conference on Machine Learning. PMLR, pp. 17061–17084.

Koide, T., Fukushi, N., Nakano, H., Chiba, D., 2024. Chatspamdetector: Leveraging large language models for effective phishing email detection. ArXiv preprint arXiv:2402.18093.

Krombholz, K., Hobel, H., Huber, M., Weippl, E., 2015. Advanced social engineering attacks. J. Inf. Secur. Appl. 22, 113–122.

Kumar, P., 2024. Adversarial attacks and defenses for large language models (LLMs): methods, frameworks & challenges. Int. J. Multimed. Inf. Retr. 13 (3), 26.

Kumar, V., Srivastava, P., Dwivedi, A., Budhiraja, I., Ghosh, D., Goyal, V., Arora, R., 2023. Large-language-models (LLM)-based AI chatbots: Architecture, in-depth analysis and their performance evaluation. In: International Conference on Recent Trends in Image Processing and Pattern Recognition. Springer, pp. 237–249.

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., et al., 2024. Voicebox: Text-guided multilingual universal speech generation at scale. Adv. Neural Inf. Process. Syst. 36.

Lee, M., Park, E., 2023. Real-time Korean voice phishing detection based on machine learning approaches. J. Ambient. Intell. Humaniz. Comput. 14 (7), 8173–8184.

Levitan, S.I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., Rosenberg, A., 2015. Cross-cultural production and detection of deception from speech. In: Proc. WMDD. ACM, New York, NY, USA, pp. 1–8.

Li, L., Lu, T., Ma, X., Yuan, M., Wan, D., 2023. Voice deepfake detection using the self-supervised pre-training model HuBERT. Appl. Sci. 13 (14), 8488.

Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., et al., 2023a. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. IEEE/ACM Trans. Audio Speech Lang. Process. 31, 2507–2522.

Liu, C., Zhang, J., Fang, H., Ma, Z., Zhang, W., Yu, N., 2023b. Dear: A deep-learning-based audio re-recording resilient watermarking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11. pp. 13201–13209.

Liu, P., Zhang, L., Gulla, J.A., 2023c. Pre-train, prompt, and recommendation: A comprehensive survey of language modeling paradigm adaptations in recommender systems. Trans. Assoc. Comput. Linguist. 11, 1553–1571.

Liu, G., Zhang, Y., Lei, Y., Chen, Y., Wang, R., Li, Z., Xie, L., 2023d. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. ArXiv preprint arXiv:2305.19522.

Liu, T., Zhang, Y., Zhao, Z., Dong, Y., Meng, G., Chen, K., 2024. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. ArXiv preprint arXiv:2402.18104.

Lloyd, E.P., Deska, J.C., Hugenberg, K., McConnell, A.R., Humphrey, B.T., Kunstman, J.W., 2019. Miami university deception detection database. Behav. Res. Methods 51, 429–439.

Ma, X., Zhang, S., Huang, S., Gao, J., Hu, Y., He, L., 2023. How to boost anti-spoofing with x-vectors. In: 2022 IEEE Spoken Language Technology Workshop. SLT, IEEE, pp. 593–598.

Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., Isaac, W., 2024. Generative AI misuse: A taxonomy of tactics and insights from real-world data. ArXiv preprint arXiv:2406.13843.

Mayew, W.J., Venkatachalam, M., 2012. The power of voice: Managerial affective states and future firm performance. J. Financ. 67 (1), 1–43.

Mihalache, S., Pop, G., Burileanu, D., 2019. Introducing the rodecar database for deceptive speech detection. In: 2019 International Conference on Speech Technology and Human-Computer Dialogue. SpeD, pp. 1–6.

Montañez, R., Golob, E., Xu, S., 2020. Human cognition through the lens of social engineering cyberattacks. Front. Psychol. 11, 1755. http://dx.doi.org/10.3389/fpsyg.2020.01755, URL: https://www.frontiersin.org/article/10.3389/fpsyg.2020.01755/full.

Nahari, G., Ashkenazi, T., Fisher, R.P., Granhag, P.-A., Hershkowitz, I., Masip, J., Meijer, E.H., Nisin, Z., Sarid, N., Taylor, P.J., et al., 2019. 'Language of lies': Urgent issues and prospects in verbal lie detection research. Leg. Criminol. Psychol. 24 (1), 1–23.

Nasri, H., Ouarda, W., Alimi, A.M., 2016. Relidss: Novel lie detection system from speech signal. In: 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications. AICCSA, pp. 1–8.

Nicholson, J., Coventry, L., Briggs, P., 2019. "If it's important it will be a headline": Cybersecurity information seeking in older adults. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–11. http://dx.doi.org/10.1145/3290605.3300579, URL: https://doi.org/10.1145/3290605.3300579.

O'Keefe, D.J., 2013. The elaboration likelihood model. In: The Sage Handbook of Persuasion: Developments in Theory and Practice. Sage Thousand Oaks, CA, pp. 137–149.

Pang, K., Qi, T., Wu, C., Bai, M., 2024. Adaptive and robust watermark against model extraction attack. ArXiv preprint arXiv:2405.02365.

Pavlović, K., Kovačević, S., Djurović, I., Wojciechowski, A., 2022. Robust speech watermarking by a jointly trained embedder and detector using a DNN. Digit. Signal Process. 122, 103381.

Peng, B., Li, C., He, P., Galley, M., Gao, J., 2023. Instruction tuning with gpt-4. ArXiv preprint arXiv:2304.03277.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M., 2015a. Deception detection using real-life trial data. In: Proc. ICMI. pp. 59–66.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Xiao, Y., Linton, C., Burzo, M., 2015b. Verbal and nonverbal clues for real-life deception detection. In: Proc. EMNLP. pp. 2336–2346.

Pérez-Rosas, V., Mihalcea, R., Narvaez, A., Burzo, M., 2014. A multimodal dataset for deception detection. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC'14, European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 3118–3122.

Petty, R.E., Cacioppo, J.T., Petty, R.E., Cacioppo, J.T., 1986. The Elaboration Likelihood Model of Persuasion. Springer.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M., 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In: International Conference on Machine Learning. PMLR, pp. 8599–8608.

Prange, S., Knierim, P., Knoll, G., Dietz, F., Luca, A.D., Alt, F., 2024. 'I do (not) need that Feature!" – understanding Users' awareness and control of privacy permissions on android smartphones. In: Twentieth Symposium on Usable Privacy and Security. SOUPS 2024, USENIX Association, Philadelphia, PA, pp. 453–472, URL: https://www.usenix.org/conference/soups2024/presentation/prange.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. 1 (2), 3, ArXiv preprint arXiv:2204.06125.

Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A., 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. ArXiv preprint arXiv:2402.07927.

Salvi, F., Ribeiro, M.H., Gallotti, R., West, R., 2024. On the conversational persuasiveness of large language models: A randomized controlled trial. ArXiv preprint arXiv:2403.14380.

San Roman, R., Fernandez, P., Elsahar, H., Défossez, A., Furon, T., Tran, T., 2024. Proactive detection of voice cloning with localized watermarking. In: International Conference on Machine Learning. vol. 235.

Scherer, K.R., Feldstein, S., Bond, R.N., Rosenthal, R., 1985. Vocal cues to deception: A comparative channel approach. J. Psycholinguist. Res. 14, 409–425.
Schuller, B., 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61 (5), 90–99.
Schuller, B.W., Amiriparian, S., Batliner, A., Gebhard, A., Gerzcuk, M., Karas, V., Kathan, A., Seizer, L., Löchner, J., 2022. Computational charisma–a brick by brick blueprint for building charismatic artificial intelligence. ArXiv preprint arXiv:2301.00142.
Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J.K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., Evanini, K., 2016. The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language. In: Proc. INTERSPEECH. pp. 2001–2005.
Soldner, F., Pérez-Rosas, V., Mihalcea, R., 2019. Box of Lies: Multimodal deception detection in dialogues. In: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 1768–1777.
Sorokowski, P., Puts, D., Johnson, J., Żółkiewicz, O., Oleszkiewicz, A., Sorokowska, A., Kowal, M., Borkowska, B., Pisanski, K., 2019. Voice of authority: professionals lower their vocal frequencies when giving expert advice. J. Nonverbal Behav. 43, 257–269.
Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., Lin, J., 2024. Large language models for forecasting and anomaly detection: A systematic literature review. ArXiv preprint arXiv:2402.10350.
Tan, X., 2023. Neural Text-To-Speech Synthesis. Springer.
Tang, H., Zhang, X., Cheng, N., Xiao, J., Wang, J., 2024. ED-TTS: Multi-scale emotion modeling using cross-domain emotion diarization for emotional speech synthesis. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 12146–12150.
The European Parliament, 2023. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.
Tigue, C.C., Borak, D.J., O'Connor, J.J., Schandl, C., Feinberg, D.R., 2012. Voice pitch influences voting behavior. Evol. Hum. Behav. 33 (3), 210–216.
Triantafyllopoulos, A., Schuller, B.W., 2024. Expressivity and speech synthesis. ArXiv preprint arXiv:2404.19363.
Triantafyllopoulos, A., Schuller, B.W., İymen, G., Sezgin, M., He, X., et al., 2023. An overview of affective speech synthesis and conversion in the deep learning era. Proc. IEEE 111 (10), 1355–1381.
Triantafyllopoulos, A., Tsangko, I., Gebhard, A., Mesaros, A., Virtanen, T., Schuller, B., 2024. Computer audition: From task-specific machine learning to foundation models. ArXiv preprint arXiv:2407.15672.
Vance, N., Speth, J., Khan, S., Czajka, A., Bowyer, K.W., Wright, D., Flynn, P., 2022. Deception detection and remote physiological monitoring: A dataset and baseline experimental results. IEEE Trans. Biom. Behav. Identity Sci. 4 (4), 522–532.
Volkamer, M., Renaud, K., Reinheimer, B., 2016. TORPEDO: Tooltip-powered phishing email detection. In: Hoepman, J.-H., Katzenbeisser, S. (Eds.), ICT Systems Security and Privacy Protection. Springer International Publishing, Cham, pp. 161–175.
Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., Schuller, B.W., 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Trans. Pattern Anal. Mach. Intell.
Wang, Y., 2024. Survey for detecting AI-generated content. Adv. Eng. Technol. Res. 11 (1), 643.
Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al., 2023. Neural codec language models are zero-shot text to speech synthesizers. ArXiv preprint arXiv:2301.02111.
Wang, Z., Huang, Y., Song, D., Ma, L., Zhang, T., 2024. PromptCharm: Text-to-image generation through multi-modal prompting and refinement. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–21.
Wang, X., Yamagishi, J., 2021. Investigating self-supervised front ends for speech spoofing countermeasures. ArXiv preprint arXiv:2111.07725.
Wang, L., Yoshida, Y., Kawakami, Y., Nakagawa, S., 2015. Relative phase information for detecting human speech and spoofed speech. In: Interspeech.
Warnita, T., Lestari, D.P., 2017. Construction and analysis of Indonesian-interviews deception corpus. In: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment. O-COCOSDA, pp. 1–6.
Wu, Z., Xiao, X., Chng, E.S., Li, H., 2013. Synthetic speech detection using temporal modulation feature. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7234–7238.
Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D.F., Chao, L.S., 2023. A survey on llm-gernerated text detection: Necessity, methods, and future directions. ArXiv preprint arXiv:2310.14724.
Yang, Z., Jing, X., Triantafyllopoulos, A., Song, M., Aslan, I., Schuller, B.W., 2022. An overview & analysis of sequence-to-sequence emotional voice conversion. ArXiv preprint arXiv:2203.15873.
Yang, D., Liu, S., Huang, R., Weng, C., Meng, H., 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. IEEE/ACM Trans. Audio Speech Lang. Process.
Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., Zhang, Y., 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confid. Comput. 100211.
Yeo, I.-K., Kim, H.J., 2003. Modified patchwork algorithm: A novel audio watermarking scheme. IEEE Trans. Speech Audio Process. 11 (4), 381–386.
Yi, J., Tao, J., Fu, R., Yan, X., Wang, C., Wang, T., Zhang, C.Y., Zhang, X., Zhao, Y., Ren, Y., et al., 2023. Add 2023: the second audio deepfake detection challenge. ArXiv preprint arXiv:2305.13774.
Yu, Y., Liu, X., Ni, R., Yang, S., Zhao, Y., Kot, A.C., 2023. Pvass-mdd: predictive visual-audio alignment self-supervision for multimodal deepfake detection. IEEE Trans. Circuits Syst. Video Technol.
Yu, H., Tan, Z.-H., Zhang, Y., Ma, Z., Guo, J., 2017. DNN filter bank cepstral coefficients for spoofing detection. IEEE Access 5, 4779–4787.
Yuan, Z., Xiong, Z., Zeng, Y., Yu, N., Jia, R., Song, D., Li, B., 2024. Rigorllm: Resilient guardrails for large language models against undesired content. ArXiv preprint arXiv:2403.13031.
Zeinali, H., Stafylakis, T., Athanasopoulou, G., Rohdin, J., Gkinis, I., Burget, L., Černockỳ, J., et al., 2019. Detecting spoofing attacks using vgg and sincnet: but-omilia submission to asvspoof 2019 challenge. ArXiv preprint arXiv:1907.12908.
Zhang, G., Gao, M., Li, Q., Zhai, W., Jeon, G., 2024a. Multi-modal generative DeepFake detection via visual-language pretraining with gate fusion for cognitive computation. Cogn. Comput. 1–14.
Zhang, C., Jin, M., Yu, Q., Liu, C., Xue, H., Jin, X., 2024b. Goal-guided generative prompt injection attack on large language models. ArXiv preprint arXiv:2404.07234.
Zhang, Y., Li, Z., Lu, J., Hua, H., Wang, W., Zhang, P., 2023. The impact of silence on speech anti-spoofing. IEEE/ACM Trans. Audio Speech Lang. Process.
Zheng, S.Y., Becker, I., 2023. Checking, nudging or scoring? Evaluating e-mail user security tools. In: Nineteenth Symposium on Usable Privacy and Security. SOUPS 2023, pp. 57–76.
Zuckerman, M., DePaulo, B.M., Rosenthal, R., 1981. Verbal and nonverbal communication of deception. In: Advances in Experimental Social Psychology. vol. 14, Elsevier, pp. 1–59.