

## Article

# Human Detection of Voice-Cloned Speech Under GSM, VoLTE and VoIP Conditions

Jakub Warzych <sup>1</sup>, Michał Łuczyński <sup>2,\*</sup> and Janusz Klink <sup>1</sup>

<sup>1</sup> Department of Telecommunications and Teleinformatics, Faculty of Information and Telecommunication Technology, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland; 263858@student.pwr.edu.pl (J.W.); janusz.klink@pwr.edu.pl (J.K.)

<sup>2</sup> Department of Acoustics, Multimedia and Signal Processing, Faculty of Electronics, Photonics and Microsystems, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

\* Correspondence: [michal.luczynski@pwr.edu.pl](mailto:michal.luczynski@pwr.edu.pl)

## Abstract

The rapid progress of generative speech synthesis and voice-cloning technologies has enabled the creation of highly natural synthetic voices that pose a serious threat to telecommunication security. While most prior studies evaluate human ability to detect audio deepfakes using high-quality, studio-grade recordings, little is known about how real-world telecommunication channels affect perceptual detection. This study investigates the influence of three transmission scenarios—GSM (AMR-NB), VoLTE (AMR-WB), and VoIP with packet-loss modeling—on the human ability to distinguish natural speech from AI-generated speech. A custom speech corpus was developed, consisting of natural recordings from nine speakers and corresponding synthetic utterances generated using a state-of-the-art voice cloning system (ElevenLabs). All samples were processed through simulated telecommunication channels using real codec implementations. A listening test with 95 participants was conducted, involving binary classification (human vs. synthetic) and confidence ratings. Results show an overall detection accuracy of 54.8%, confirming that humans are poorly equipped to identify synthetic speech. Surprisingly, the highest accuracy was achieved for the narrowband GSM channel (63.7%), while VoLTE yielded the lowest performance (44.0%). The findings suggest that restricted bandwidth may emphasize prosodic irregularities typical of generative models, whereas high-quality channels mask synthetic artifacts, increasing susceptibility to voice spoofing. The results highlight the necessity of deploying additional security mechanisms in telecommunication systems relying on voice identity verification.

**Keywords:** human perception; voice cloning; detection; cross-channel; telecommunication; GSM; VoLTE; VoIP



Academic Editor: Michel Darmon

Received: 15 April 2026

Revised: 8 June 2026

Accepted: 11 June 2026

Published: 17 June 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

The rapid progress of generative artificial intelligence has led to a profound transformation in modern speech synthesis technologies. Contemporary text-to-speech (TTS) systems are no longer limited to producing intelligible but artificial-sounding speech. Instead, state-of-the-art neural architectures are capable of generating speech that closely resembles natural human voices in terms of timbre, prosody, and speaking style. A particularly impactful development within this domain is voice cloning, which enables the

synthesis of speech that mimics the vocal identity of a specific individual from only a short reference recording [1–3].

Recent models based on end-to-end neural architectures, including diffusion-based approaches and neural audio codecs, have reached perceptual quality comparable to or exceeding that of professionally recorded speech in controlled listening tests [1,2]. Zero-shot and few-shot paradigms further reduce the requirements for speaker-specific data, making realistic voice cloning widely accessible beyond research environments [2,3]. Furthermore, the emergence of neural codec language models has enabled high-fidelity zero-shot synthesis with even shorter prompts [4]. Commercial platforms utilizing these techniques now allow users to generate highly convincing speech using minimal input data, effectively democratizing voice cloning technology.

While these advances unlock numerous beneficial applications, such as assistive technologies, localization, and content creation, they also introduce serious security and societal risks. AI-generated and voice-cloned speech has been increasingly used in phishing attacks, impersonation scams, and other forms of voice-based social engineering. In many of these scenarios, the victim is exposed to the synthetic voice through a telecommunication channel, such as a phone call or an internet-based voice service. Because voice cues are still commonly perceived as indicators of authenticity, high-quality synthetic speech significantly challenges traditional assumptions underlying remote identity verification.

A growing body of research has examined the ability of humans to detect AI-generated speech and audio deepfakes. Recent studies consistently show that human listeners are poorly equipped to distinguish between natural and synthetic speech, often achieving detection accuracies only slightly above chance level [5–7]. Moreover, these studies reveal that listeners frequently exhibit high confidence in incorrect judgments, further amplifying the risk of successful deception [6,7].

However, an important limitation of existing research is that most perceptual evaluations are conducted using studio-quality or uncompressed recordings, devoid of typical telecommunication distortions. In contrast, real-world communication channels—such as GSM, VoLTE, and VoIP—introduce codec-dependent bandwidth limitations, compression artifacts, packet loss, jitter, and device-specific frequency responses. To ensure methodological rigor, perceptual studies in this field often draw upon established standards for subjective speech quality assessment, such as the ITU-T P.800 recommendation [8]. These degradations may either mask subtle synthetic artifacts or, conversely, emphasize prosodic inconsistencies produced by generative models.

This leads to a key open research question addressed in this study: Do degradations introduced by practical telecommunication channels make voice-cloned speech easier or harder for humans to detect?

The objective of this work is therefore to evaluate human perceptual detection of voice-cloned speech across realistic telecommunication scenarios, specifically GSM, VoLTE, and VoIP transmission conditions. A dedicated listening experiment was conducted using a custom dataset comprising natural and voice-cloned speech samples processed through simulated telecommunication channels. The main contributions of this work are as follows:

- The creation of a custom, cross-channel dataset combining natural and voice-cloned speech;
- A direct comparison of three widely used telecommunication channels (GSM, VoLTE, VoIP);
- An analysis of both detection accuracy and listener confidence;
- Empirical evidence showing that higher technical signal quality does not necessarily improve detection performance.

This study addresses a critical gap in the literature by systematically investigating human detection of voice-cloned speech under realistic telecommunication conditions, providing insights relevant to both speech technology research and telecommunication security.

## 2. Background and Related Work

### 2.1. Voice Cloning and Modern Generative Text-to-Speech

Voice cloning has emerged as a key capability of modern generative speech synthesis systems. Unlike traditional TTS approaches, such as unit selection [9] or statistical parametric synthesis [10], which typically produce speech with generic voices, voice cloning aims to reconstruct the vocal characteristics of a specific speaker.

Neural architectures such as WaveNet [11], Tacotron-based systems [12], VITS [13], and more recent diffusion-based models like NaturalSpeech 3 [2] operate in an end-to-end manner, learning complex mappings between text representations and raw audio waveforms. More recently, neural codec language models like VALL-E [4] have demonstrated the ability to clone voices using as little as a three-second enrollment recording. Zero-shot voice cloning approaches leverage speaker embeddings extracted from short reference recordings, enabling generalization to unseen voices without retraining the synthesis model [2,3]. This paradigm has become a standard component of current state-of-the-art systems. Subjective evaluations indicate that the perceptual quality of such systems approaches human recordings in controlled scenarios [1,2], significantly complicating perceptual discrimination.

### 2.2. Human Perception and Detection of Audio Deepfakes

With the increasing realism of AI-generated speech, research attention has shifted from quality assessment toward detectability. Detection capabilities are tested both through the development of machine learning models (including challenges such as ASVspoof [14]) and through listening tests. Multiple studies have examined whether humans can reliably identify synthetic speech. Barrington et al. demonstrated that participants were largely unable to detect AI-powered voice clones, even when warned about their presence [5]. Similarly, Mai et al. reported detection accuracies close to chance level in online listening experiments, with minimal improvement following short training [6].

Milewski et al. compared human listeners with neural network-based detectors and observed that while machine learning models could exploit subtle spectral cues, humans consistently relied on perceptually salient but increasingly unreliable indicators [7]. A recurring finding is the disconnect between confidence and correctness: listeners often express strong confidence in their judgments despite objectively poor performance [6,7]. Importantly, nearly all existing perceptual studies evaluate detection using clean recordings, which may not generalize to everyday communication.

### 2.3. Telecommunication Codecs and Channel Degradations

Speech transmitted through telecommunication systems undergoes substantial processing. In traditional GSM networks, narrowband codecs such as AMR-NB constrain the signal to a bandwidth of approximately 300–3400 Hz, removing higher-frequency components [15]. VoLTE employs wideband codecs (AMR-WB), extending the bandwidth to around 7 kHz and improving clarity. VoIP systems rely on packet-based transmission, introducing time-varying impairments such as packet loss and jitter. The perceived quality in such networks is often modeled using frameworks like the E-model, which accounts for the cumulative effect of equipment impairment and network conditions [16]. These degradations may be misattributed by listeners to synthesis artifacts or distract attention from genuine inconsistencies produced by generative models.

### 2.4. Speech Perception, Prosody, and Synthetic Artifacts

Human perception of speech authenticity is influenced by prosodic features such as intonation, rhythm, and stress. While modern models have overcome spectral limitations, they may still exhibit subtle irregularities in prosodic control [6,7]. Psychophysical studies

suggest these cues are crucial for subconscious judgments of naturalness. Notably, many of these cues are concentrated in lower frequency bands, implying that bandwidth-limited channels (GSM) may inadvertently emphasize features where generative models remain imperfect, while high-quality transmission (VoLTE) may bias listeners toward perceiving speech as authentic due to increased clarity.

### 2.5. Research Gap

Although prior work has clearly established the difficulty humans face in detecting voice-cloned speech, the interaction between voice cloning and realistic telecommunication channels remains underexplored. Existing studies predominantly assess detection performance under idealized acoustic conditions, neglecting the environments in which voice spoofing attacks most commonly occur.

To date, no systematic perceptual study has compared human detection of voice-cloned speech across GSM, VoLTE, and VoIP channels, nor analyzed how channel-dependent degradations affect both detection accuracy and listener confidence. Addressing this gap is essential for understanding real-world vulnerability to voice-based deception, which directly motivates the experimental investigation presented in this work.

## 3. Materials and Methods

### 3.1. Speech Dataset

#### 3.1.1. Natural Speech Recordings

A dedicated dataset of natural speech recordings was created specifically for this study to ensure full control over acoustic quality, speaker characteristics, and recording conditions. The dataset comprises recordings from nine native Polish speakers (seven male and two female), aged between 20 and 24 years. All speakers were fluent, without known speech or hearing impairments.

Recordings were conducted in an acoustically treated room with a short reverberation time ( $RT \approx 0.2$  s) to minimize room coloration and environmental noise. Speech was captured using an AKG C2000B condenser microphone (AKG Acoustics GmbH, Vienna, Austria), connected to an Audient EVO4 audio interface (Audient Ltd., Andover, Hampshire, UK). Each speaker read a predefined speech script containing neutral, context-independent utterances designed to resemble realistic conversational speech typically occurring during telephone communication. The content avoided named entities and emotionally loaded phrases in order to reduce semantic bias.

Audio signals were recorded using Reaper software (Cockos Inc., New York, NY, USA), version 7.x at a sampling rate of 44.1 kHz and 16-bit resolution, and saved in uncompressed WAV format. Each recording session resulted in several short speech segments, from which final test samples of 8–12 s duration were selected.

#### 3.1.2. Synthetic Speech Generation (Voice Cloning)

Synthetic speech samples were generated using a state-of-the-art commercial voice cloning platform (ElevenLabs) (ElevenLabs Inc., New York, NY, USA), employing a few-shot voice cloning paradigm. For each of the nine speakers, a speaker-specific voice model was created using a short reference recording extracted from the natural speech dataset.

Speech synthesis was performed through the ElevenLabs application programming interface (API) to ensure access to the highest available output quality. All synthetic utterances were generated in linear PCM format with a sampling rate of 24 kHz, avoiding lossy compression at the generation stage. Each synthetic sample corresponded semantically and stylistically to one natural utterance, while remaining distinct to prevent memorization effects.

All generated samples were verified through manual listening to remove cases with pronunciation errors, audible artifacts, or synthesis failures. The final dataset consisted of nine unique natural speech samples and nine corresponding voice-cloned samples, all matched in duration and linguistic complexity.

### *3.2. Telecommunication Channel Simulation*

To evaluate perceptual detection under realistic conditions, all speech samples were processed through three representative telecommunication scenarios using offline simulation. Channel modeling was performed using FFmpeg (FFmpeg Developers, open-source project, global), version 7.x, relying on standardized codec implementations and signal processing chains.

#### *3.2.1. GSM Channel*

The GSM scenario represents traditional circuit-switched mobile telephony. Speech samples were processed using the AMR-NB codec (Adaptive Multi-Rate Narrowband) codec standardized by 3GPP (3rd Generation Partnership Project, global standardization body), implemented via FFmpeg libraries (version as above) at a fixed bitrate of 12.2 kb/s, corresponding to the highest quality configuration commonly deployed in GSM networks. The signal bandwidth was limited to approximately 300–3400 Hz, emulating narrow-band transmission.

#### *3.2.2. VoLTE Channel*

The VoLTE scenario models modern IP-based mobile telephony with enhanced audio quality. Samples were encoded and decoded using the AMR-WB codec (Adaptive Multi-Rate Wideband) codec standardized by 3GPP (3rd Generation Partnership Project, global standardization body), implemented via FFmpeg libraries (version as above) at 12.65 kb/s, providing a wideband frequency range of approximately 50–7000 Hz. This configuration reflects typical Voice over LTE deployments and represents the highest perceived quality among the tested scenarios.

#### *3.2.3. VoIP Channel*

The VoIP condition was designed to emulate internet-based voice communication, where speech quality is influenced not only by codec performance but also by network impairments. Instead of fixed codec compression, the VoIP scenario introduced packet-loss-like disruptions, temporal jitter, and background noise. Random short dropouts were applied to simulate packet loss, combined with slight temporal irregularities and low-level pink noise to approximate network-induced disturbances.

### *3.3. Signal Conditioning and Normalization*

In order to simulate realistic end-user recording conditions, a microphone frequency response model consistent with mobile handset specifications was applied before codec processing. The frequency shaping was based on tolerance masks defined in ETSI TS 26.131, separately for narrowband (GSM) and wideband (VoLTE/VoIP) operation. This step ensured that the spectral characteristics of the speech matched those of typical mobile device microphones.

All output signals were then resampled to a common sampling rate of 16 kHz, converted to mono, and loudness-normalized to  $-1$  dBFS. This normalization prevented loudness cues from influencing listener judgments while preserving natural dynamics. No additional noise reduction, compression, or enhancement processing was applied.

### 3.4. Subjective Listening Test

#### 3.4.1. Experimental Design

A subjective listening experiment was conducted using a binary classification task. Each participant was presented with 18 speech samples, consisting of:

- 9 natural speech samples,
- 9 voice-cloned speech samples,
- evenly distributed across the three transmission conditions (GSM, VoLTE, VoIP).

Samples were presented individually in randomized order. After listening to each sample, participants were asked to indicate whether they believed the voice originated from a human speaker or artificial intelligence, with an optional “unsure” response. Subsequently, participants rated their confidence level on a five-point Likert scale, ranging from complete uncertainty (1) to full confidence (5).

#### 3.4.2. Participants

A total of 95 participants took part in the experiment. Recruitment was conducted via two channels:

- Online recruitment through the Prolific platform, providing a diverse sample of listeners using personal equipment,
- a local expert group consisting of members of an academic audio engineering student society, who participated under controlled listening conditions.

Participants represented a broad range of ages and levels of familiarity with artificial intelligence technologies. Prior to the listening task, demographic data and self-reported experience with AI and voice cloning were collected via questionnaire.

### 3.5. Statistical Analysis

Detection performance was quantified as classification accuracy across conditions. In addition to descriptive statistics, inferential analysis was conducted to compare detection performance between telecommunication channels.

A one-way repeated-measures ANOVA was applied to assess the effect of channel type (GSM, VoLTE, VoIP) on detection accuracy. Where assumptions of normality were violated, a non-parametric alternative (Kruskal–Wallis test) was employed. Post hoc pairwise comparisons were performed using Tukey’s HSD (for ANOVA) or Dunn’s test with Bonferroni correction (for non-parametric analysis). Listener confidence data were analyzed analogously.

Significance was evaluated at a threshold of  $p < 0.05$ . In addition to the accuracy-based analysis, a basic Signal Detection Theory (SDT) analysis was performed to estimate sensitivity ( $d'$ ) and decision criterion ( $c$ ). Because the listening task included a third response option (“unsure”), SDT metrics were computed using only binary responses (AI vs. Human), while “unsure” responses were excluded from the SDT calculation. Hit rates and false alarm rates were calculated separately for GSM, VoIP, and VoLTE conditions, and  $d'$  and  $c$  were derived from the corrected proportions using the standard normal inverse cumulative distribution.

## 4. Experimental Procedure

### 4.1. Overall Experimental Design

The listening experiment was designed to evaluate human ability to distinguish natural speech from voice-cloned speech under realistic telecommunication conditions. The study followed a between-conditions, within-participant design, in which each participant

evaluated speech samples processed through all three simulated transmission scenarios: GSM, VoLTE, and VoIP.

To balance experimental control with ecological validity, understood here as realistic end-user listening conditions, the study was conducted in two complementary settings:

- an online listening experiment, reflecting typical end-user listening environments,
- a controlled laboratory experiment, conducted with a smaller expert group under standardized acoustic conditions.

Both test versions followed the same stimulus set, task structure, and questionnaire design, ensuring direct comparability of results.

Alongside the perceptual experiments, an objective acoustic analysis was conducted on the complete speech stimuli corpus to quantify the signal-level differences between natural and synthetic voices. This technical characterization establishes the physical baseline of the audio signals before examining how these acoustic profiles modulate human classification performance under different network conditions.

#### 4.2. Online Listening Test

The majority of participants completed the experiment remotely using an online survey and audio playback interface. Participants were recruited via the Prolific platform, which enables selective recruitment and quality control while allowing listeners to use their personal playback equipment (headphones or loudspeakers).

Before beginning the listening task, participants were instructed to:

- perform the test in a quiet environment,
- use headphones if possible,
- avoid multitasking during the experiment.

Although full control over listening conditions was not possible in the online setting, this approach reflects realistic everyday communication scenarios in which voice-based deception is most likely to occur.

Audio samples were presented sequentially, and playback was restricted to prevent repeated listening to the same sample. The total duration of the experiment did not exceed 10 min, minimizing fatigue effects and maintaining participant attention.

#### 4.3. Controlled Listening Test

In addition to the online experiment, a subset of participants formed a local expert group, consisting of members of an academic audio engineering and acoustics student society. This group completed the listening test in a controlled acoustic environment, using identical playback equipment and standardized exposure conditions.

The controlled test was conducted in small groups of up to seven participants. Samples were presented in a fixed temporal sequence, with:

- a short preparation signal preceding each sample,
- a predefined response window,
- no possibility to pause, replay, or skip samples.

This procedure ensured consistent exposure across participants and eliminated variability related to individual playback devices or listening environments. The inclusion of this group enabled partial separation of perceptual effects from environmental confounds.

#### 4.4. Questionnaire Structure

The experimental procedure was implemented as a three-part questionnaire:

Part 1: Information and Consent

Participants were informed about the voluntary and anonymous nature of the experiment. No information regarding the proportion of synthetic or natural samples was disclosed to avoid expectation bias.

#### Part 2: Demographic and Background Information

Participants provided information regarding:

- age group,
- gender (optional),
- frequency of interaction with AI-based tools,
- prior familiarity with voice cloning or synthetic speech technologies,
- type of playback device used during the experiment.

This information was later used to analyze potential correlations between listener characteristics and detection performance.

#### Part 3: Listening Task

In the listening task, each participant evaluated 18 speech samples, consisting of:

- 9 natural recordings,
- 9 voice-cloned recordings,
- evenly distributed across GSM, VoLTE, and VoIP conditions.

After each sample, participants performed two actions:

1. Binary classification, indicating whether the voice originated from a human or artificial intelligence (with an optional “unsure” response),
2. Confidence rating, expressing how confident they were in their decision using a five-point Likert scale.

#### 4.5. Participant Characteristics

A total of 95 participants completed the experiment. The sample included listeners from multiple age groups, with the majority between 18 and 35 years of age, reflecting the dominant demographic of users of modern telecommunication services.

With respect to technological familiarity:

- approximately two-thirds of participants reported frequent interaction with AI-based tools (at least once per week),
- a smaller subset reported prior experience with voice cloning or speech synthesis technologies,
- the remaining participants had limited or no direct exposure to synthetic speech systems.

This diversity allowed for comparative analysis across demographic and experiential factors, including age, AI familiarity, and technical background.

#### 4.6. Data Quality and Ethical Considerations

To ensure data quality, online responses were screened for implausibly short completion times and incomplete questionnaires. No personally identifiable information was collected. All participants provided informed consent, and the experiment posed no foreseeable risk.

The combination of online and controlled experiments provided a robust methodological framework, enabling investigation of perceptual detection under both realistic and standardized listening conditions.

#### 4.7. Automatic Deepfake Detection Framework

An automatic speech deepfake detection baseline was introduced into the experimental pipeline to provide a technical benchmark for the human perceptual study. Two distinct classification architectures were implemented to evaluate different tiers of feature extrac-

tion. The first architecture, establishing a classical machine learning baseline, utilizes 40-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) paired with a Random Forest (RF) classifier configured with 100 estimators. The second architecture comprises an advanced Self-Supervised Learning (SSL) pipeline utilizing feature embeddings extracted from the pretrained Wav2Vec2 architecture (wav2vec2-base), which are subsequently classified via a Linear Support Vector Classifier (LinearSVC).

To ensure a strict, speaker-independent evaluation and eliminate potential speaker identity bias or data leakage, a Leave-One-Subject-Out Cross-Validation (LOSO-CV) protocol was adopted. Since the evaluation dataset comprises nine unique speakers, the cross-validation was executed in nine distinct folds. In each cross-validation fold, all training samples, encompassing both authentic and synthesized clones belonging to a specific speaker, were completely omitted from the training phase and reserved exclusively for the evaluation set.

Classification was performed on 1 s audio chunks extracted from the continuous speech signals, with the final file-level classification decision determined via majority voting using a calibrated threshold of 65% for the spoofed class. This conservative threshold was adopted to minimize false-positive detections and ensure that a file is classified as a deepfake only when synthesis artifacts are consistently present across multiple segments of the utterance. To address the domain mismatch introduced by non-linear telecommunication channels, both models were evaluated under two training scenarios: a clean scenario, where models were trained strictly on raw, near-studio quality data, and an augmented scenario, which utilized data augmented with simulated GSM, VoIP, and VoLTE channel distortions.

## 5. Results

### 5.1. Objective Acoustic Analysis of Speech Stimuli

Prior to evaluating data from the listening tests, an objective acoustic analysis was performed to isolate the physical parameters distinguishing natural human speech from AI-generated voice clones. Table 1 aggregates the calculated statistical metrics across the entire stimulus database, presenting a primary contrast between Human and AI speech, further categorized by the simulated transmission channels (GSM, VoLTE, and VoIP). The analyzed parameters include Fundamental Frequency mean and variability (F0 Mean and F0 variability), Jitter, Shimmer, Harmonics-to-Noise Ratio (HNR), Spectral Centroid, and Root Mean Square Energy Modulation (RMS Energy Mod).

**Table 1.** Acoustic metrics for Human and AI speech samples across transmission channels.

| Channel / Source | F0 Mean [Hz]   | F0 Variability [Hz] | Jitter [%]  | Shimmer [%]  | HNR [dB]     | Spectral Centroid [Hz] | RMS Energy Mod |
|------------------|----------------|---------------------|-------------|--------------|--------------|------------------------|----------------|
| GSM/AI           | 172.74 ± 77.39 | 50.25 ± 12.40       | 2.78 ± 0.55 | 10.67 ± 1.23 | 12.21 ± 2.63 | 1122.00 ± 52.96        | 0.060 ± 0.007  |
| GSM/Human        | 176.17 ± 72.32 | 40.59 ± 21.18       | 2.81 ± 0.82 | 11.21 ± 2.32 | 12.87 ± 2.71 | 1383.14 ± 105.90       | 0.040 ± 0.010  |
| VoIP/AI          | 177.72 ± 78.26 | 44.18 ± 10.31       | 2.39 ± 0.60 | 10.51 ± 2.88 | 11.29 ± 3.46 | 1832.69 ± 64.14        | 0.052 ± 0.018  |
| VoIP/Human       | 154.73 ± 76.76 | 33.39 ± 26.93       | 2.62 ± 0.34 | 11.47 ± 0.43 | 11.74 ± 0.54 | 1798.69 ± 208.96       | 0.061 ± 0.016  |
| VoLTE/AI         | 111.86 ± 11.30 | 29.77 ± 14.15       | 2.35 ± 0.56 | 11.72 ± 0.85 | 12.07 ± 1.37 | 1545.94 ± 93.80        | 0.061 ± 0.008  |
| VoLTE/Human      | 121.35 ± 20.80 | 32.99 ± 20.72       | 3.90 ± 0.89 | 12.98 ± 0.76 | 10.71 ± 1.14 | 1439.33 ± 118.01       | 0.054 ± 0.003  |

All values are presented as Mean ± Standard Deviation (SD).

The data compiled in Table 1 reveals distinct acoustic trends between the two signal sources, which remain present regardless of channel-specific compression. The most notable differences occur within the fundamental frequency F0 profiles and short-term micro-perturbations. Figures 1–3 present the box plot distributions for F0 variability, Jitter, and Spectral Centroid across the entire database to illustrate these structural divergences.

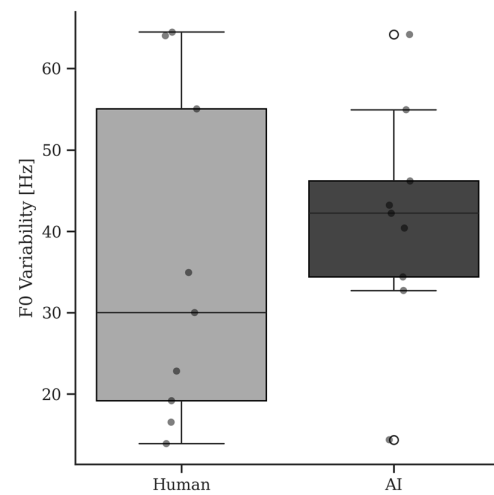


**Figure 1.** Distribution of fundamental frequency variability across human and AI voice sources. Filled dots represent individual observations, open circles indicate outliers.

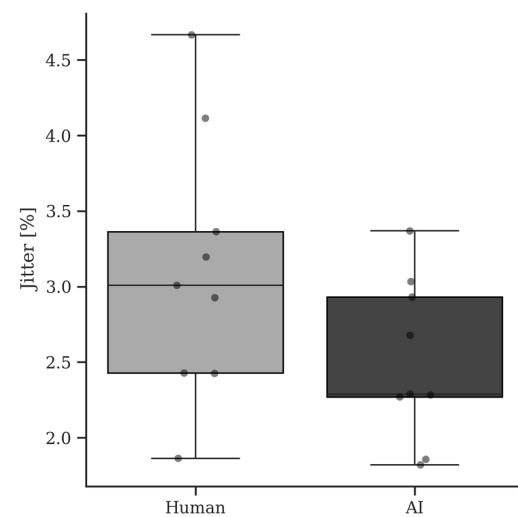


**Figure 2.** Distribution of jitter across human and AI voice sources. Filled dots represent individual observations.

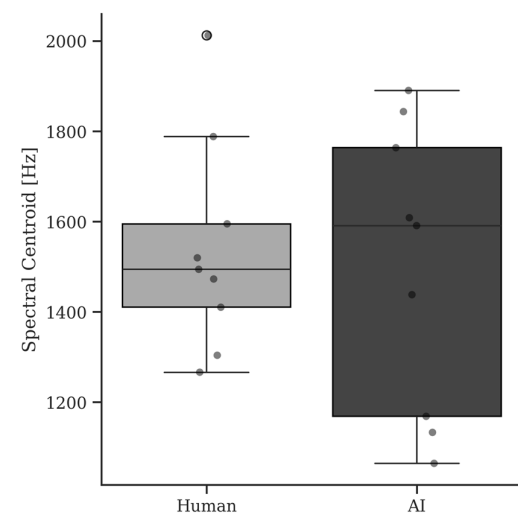


**Figure 3.** Distribution of spectral centroid metrics across human and AI voice sources. Filled dots represent individual observations, while open circle indicates outliers.

As shown in Figure 1, the overall F0 variability of the speech corpus exhibits different distribution profiles for human and synthetic speech. The human samples are characterized by a wide interquartile range (IQR) with a lower median value near 30 Hz, representing a broad and dynamic distribution of pitch variance across different speakers. Conversely, the AI samples demonstrate a much tighter, more compressed IQR centered around a higher median of approximately 42 Hz, indicating a highly consistent and less variable prosodic generation across the database.

The short-term periodic micro-variability captured by the Jitter metric (Figure 2) further clarifies this dynamic difference. Natural human speech exhibits a higher median Jitter value approximately 3.0% and a broader statistical spread, extending up to nearly 4.7%. In contrast, the synthetic AI recordings show a distinctly compressed distribution with a lower median value of around 2.3%, confirming a lower rate of cycle-to-cycle pitch micro-perturbations in the generative speech models.

Figure 3 illustrates the distribution of the Spectral Centroid across the database. Natural human samples exhibit a compact distribution concentrated between 1400 Hz and 1600 Hz with a narrow IQR. Meanwhile, the AI-generated samples present a vastly wider distribution range, spanning from below 1200 Hz to nearly 1800 Hz, which points to a higher degree of spectral dispersion and variable high-frequency energy placement in synthetic speech.

To examine how these statistical characteristics manifest in temporal signal trajectories, Figures 4 and 5 display comparative spectrograms with overlaid black F0 contours for an identical phrase produced by both sources. The human recording (Figure 4) displays characteristic vocal micro-variations, visible as continuous pitch modulations and dynamic formant transitions, particularly between the 0.0 s and 0.6 s intervals.

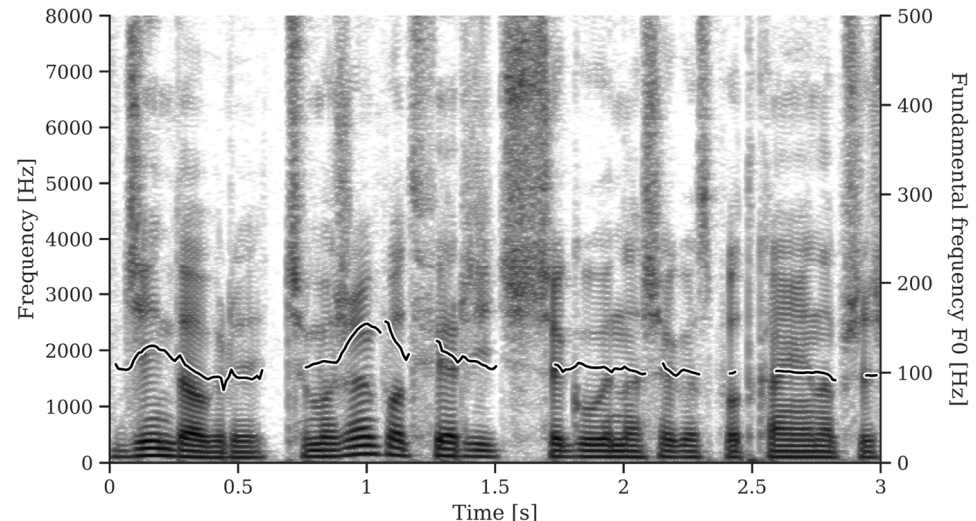


**Figure 4.** Spectrogram and F0 contour tracking for the natural human voice source.

In contrast, the AI-generated equivalent (Figure 5) exhibits distinct structural anomalies. The overlaid F0 line displays artificial smoothing and an overall rigid pitch trajectory, lacking the fine micro-perturbations seen in the human track. Furthermore, the synthetic voice contains prominent high-frequency artifacts, characterized by isolated, disconnected pitch tracking segments appearing around 7000 Hz (visible between 1.5 s and 2.5 s). These objective physical discrepancies provide a foundational baseline for interpreting the human perceptual detection results detailed in the following sections.

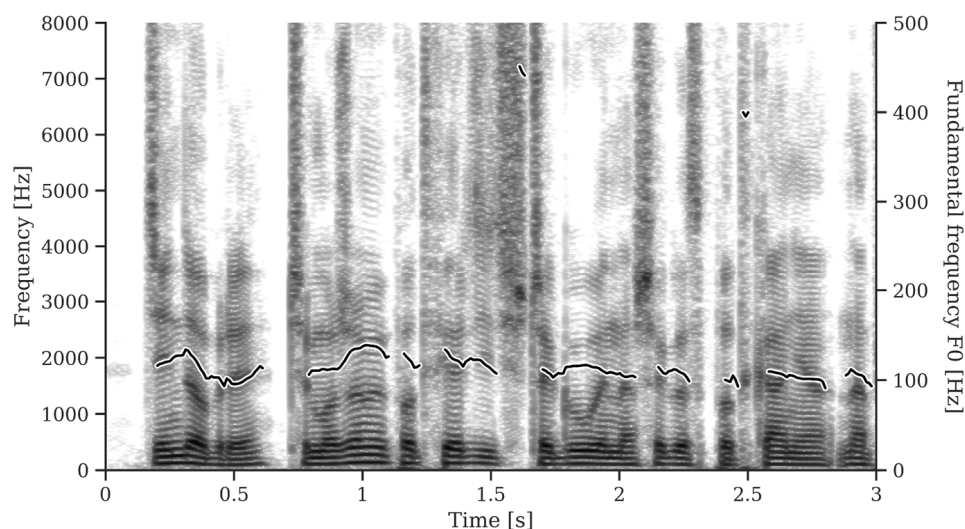


**Figure 5.** Spectrogram and F0 contour tracking for the AI-generated voice source.

### 5.2. Overall Detection Performance

The descriptive statistics of the experimental results, involving 95 participants and 1710 individual judgments, are presented in Table 2.

Statistical analysis of the results ( $N = 95$ ) showed that participants performed significantly above chance level. The mean accuracy was 54.88% (SD = 11.36), and the 95% confidence interval for the mean ranged from 52.57 to 57.20%, entirely above the chance level of 50%. The median accuracy was 56%, further indicating that a typical participant performed better than random guessing. The distribution of scores was approximately normal, with a slight negative skew ( $-0.37$ ) and near-zero kurtosis ( $-0.20$ ), suggesting the presence of a few lower scores but an overall predominance of medium and high values. Variability was moderate (IQR = 17; MAD = 6), and the observed range of scores (28–78) did not indicate the presence of outliers. Figure 6 illustrates the distribution of individual accuracy scores.

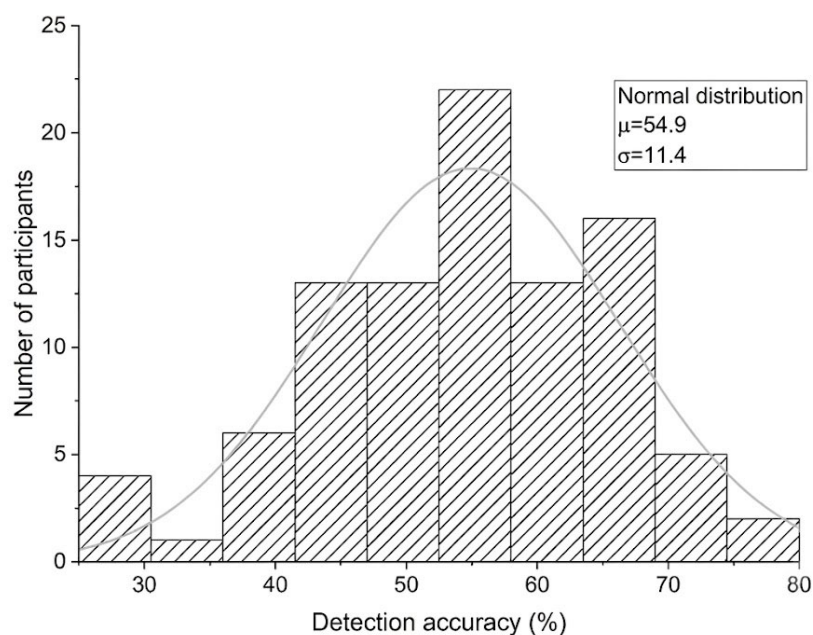


**Figure 6.** Distribution of individual detection accuracy across the participant pool ( $n = 95$ ). The histogram shows the frequency distribution, while the solid line represents a fitted normal distribution with mean  $\mu = 54.9$  and standard deviation  $\sigma = 11.4$ .

**Table 2.** Descriptive statistics of the experimental results.

| N Total | Mean | Std Dev. | Lower 95% of Mean | Upper 95% of Mean | Median | Skewness | Kurtosis | IQR <sup>(1)</sup> | MAD <sup>(2)</sup> | Min | Max |
|---------|------|----------|-------------------|-------------------|--------|----------|----------|--------------------|--------------------|-----|-----|
| 95      | 54.9 | 11.4     | 52.6              | 57.2              | 56     | −0.37    | −0.20    | 17                 | 6                  | 28  | 78  |

<sup>(1)</sup> IQR—Interquartile Range (Q3–Q1); <sup>(2)</sup> MAD—Mean Absolute Deviation.

### 5.3. Natural vs. Voice-Cloned Speech Discrimination

A subtle difference in performance was observed depending on the nature of the stimulus. Listeners were slightly more successful in identifying natural speech (mean accuracy: 57.3%) than in correctly flagging voice-cloned samples (mean accuracy: 52.3%). This 5-percentage-point gap suggests that modern voice-cloned speech is more likely to be perceived as natural than degraded natural speech is to be recognized as authentic. However, the balance of responses labeled as “Human” and “AI” indicates that no significant systematic response bias was present.

### 5.4. The Influence of Telecommunication Channels

The most significant finding of this study is the inverse relationship between technical audio fidelity and detection accuracy. As shown in Figure 7 and Table 3, transmission conditions fundamentally altered the listeners’ ability to distinguish between voices.

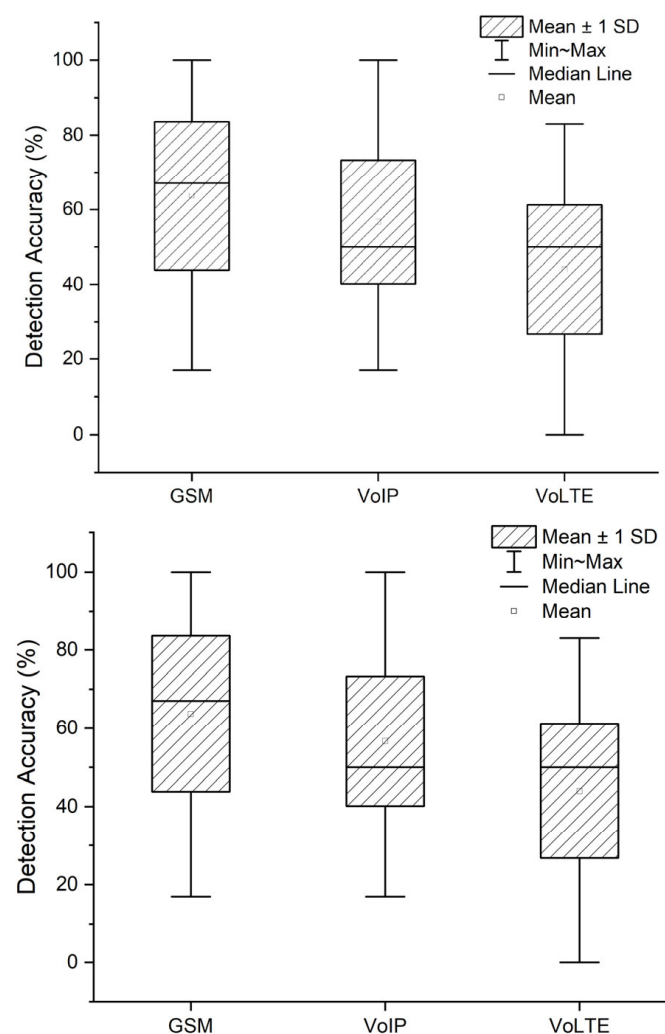
**Figure 7.** Mean detection accuracy across GSM, VoIP, and VoLTE conditions (Mean +/− SD).

**Table 3.** Detection performance and confidence metrics across transmission scenarios.

| Channel             | Accuracy [%] | Mean Confidence (1–5) | Accuracy/Confidence Gap |
|---------------------|--------------|-----------------------|-------------------------|
| GSM (Narrowband)    | 63.7         | 3.84                  | +25.3                   |
| VoIP (Packet-based) | 56.7         | 3.75                  | +19.2                   |
| VoLTE (Wideband)    | 44.0         | 3.92                  | −4.80                   |

- GSM (Narrowband): Participants achieved the highest detection accuracy (63.7%). The aggressive bandwidth limitation (300–3400 Hz) and AMR-NB compression artifacts likely emphasized mid-range prosodic inconsistencies.
- VoIP (Internet): Accuracy dropped to 56.7%, where packet-based distortions and jitter may have masked subtle synthesis cues.
- VoLTE (Wideband): Participants performed poorest in the high-quality condition (44.0%), falling significantly below chance level.

This pattern is consistent with a quality-driven tendency to associate higher spectral clarity with greater authenticity: listeners tend to subconsciously equate higher spectral clarity and wider frequency response (50–7000 Hz) with signal legitimacy. In VoLTE, the lack of typical “telephony noise” seems to lull listeners into a false sense of trust, making them more susceptible to deception.

#### 5.5. Listener Confidence and Miscalibration

Participants reported relatively high confidence in their judgments, with a mean rating of 3.77 on a 5-point scale. However, an analysis of the relationship between confidence and objective accuracy reveals a profound miscalibration of human judgment.

As summarized in Table 3, the VoLTE condition, which yielded the lowest accuracy (44.0%), elicited the highest mean confidence (3.92). This confirms that listeners are not only more likely to be deceived in high-quality channels but are also more certain of their incorrect judgments.

To further characterize the perceptual mechanisms underlying these results, a basic Signal Detection Theory (SDT) analysis was conducted for the three transmission channels. The computed metrics included sensitivity ( $d'$ ), reflecting the listener’s ability to discriminate between natural and voice-cloned speech, and decision criterion ( $c$ ), reflecting response bias toward either “Human” or “AI” judgments. Because the experimental task included an additional “unsure” category, SDT metrics were computed only from binary responses, with “unsure” responses excluded from the analysis.

Table 4 summarizes the SDT results. GSM yielded the highest sensitivity ( $d' = 0.88$ ), indicating the strongest discriminability between voice sources, accompanied by only a weak positive criterion ( $c = 0.16$ ). VoIP showed lower sensitivity ( $d' = 0.60$ ) and a negative decision criterion ( $c = -0.43$ ), indicating a bias toward “AI” responses. In contrast, VoLTE produced the weakest discriminability ( $d' = -0.20$ ) and the strongest positive criterion ( $c = 0.45$ ), indicating a bias toward “Human” responses.

**Table 4.** Signal Detection Theory metrics across transmission channels.

| Channel             | Sensitivity Index ( $d'$ ) | Decision Criterion ( $c$ ) |
|---------------------|----------------------------|----------------------------|
| GSM (Narrowband)    | 0.88                       | 0.16                       |
| VoIP (Packet-based) | 0.60                       | −0.43                      |
| VoLTE (Wideband)    | −0.20                      | 0.45                       |

Figure 8 shows detection performance across transmission channels and voice sources (voice cloning—AI or human—natural voice).

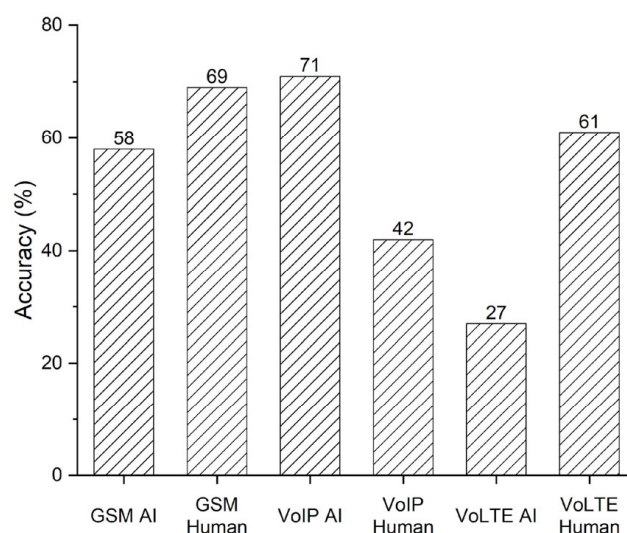


**Figure 8.** Detection performance across transmission channels and voice sources.

As Figure 8 shows, the obtained detection accuracy values differ depending on the type of voice generation source (AI or human) and the transmission channel used. The largest differences were observed for VoIP and VoLTE transmissions. In the case of VoLTE transmission, the detection accuracy rate for human-generated voice (61%) was more than twice as high as for AI-generated voice (27%). In contrast, the opposite relationship was observed for VoIP transmissions—here, detection accuracy for AI-generated voice was 71%, while for human voice it was only 42%. In the case of the GSM channel, the detection accuracy values for both types of generated voice samples were more similar and were 58% for AI-generated voice and 69% for Human voice, respectively. Descriptive statistics of the experimental results, taking into account the division into different telecommunication channels and the signal source (human/AI), are shown in Table 5.

**Table 5.** Descriptive statistics taking into account the type of channel and signal source.

| Channel / Source | N Total | Mean | Std Dev. | Lower 95% of Mean | Upper 95% of Mean | Median | Skewness | Kurtosis | IQR <sup>(1)</sup> | MAD <sup>(2)</sup> |
|------------------|---------|------|----------|-------------------|-------------------|--------|----------|----------|--------------------|--------------------|
| GSM/AI           | 95      | 0.58 | 0.28     | 0.53              | 0.64              | 0.67   | −0.38    | −0.29    | 0.33               | 0.33               |
| GSM/Human        | 95      | 0.69 | 0.28     | 0.63              | 0.75              | 0.67   | −0.37    | −0.87    | 0.67               | 0.33               |
| VoIP/AI          | 95      | 0.71 | 0.28     | 0.65              | 0.80              | 0.67   | −0.70    | −0.20    | 0.33               | 0.33               |
| VoIP/Human       | 95      | 0.42 | 0.39     | 0.34              | 0.50              | 0.33   | 0.35     | −1.38    | 0.67               | 0.33               |
| VoLTE/AI         | 95      | 0.27 | 0.30     | 0.21              | 0.33              | 0.33   | 0.92     | 0.15     | 0.33               | 0.33               |
| VoLTE/Human      | 95      | 0.61 | 0.25     | 0.56              | 0.66              | 0.67   | −0.17    | −0.27    | 0.33               | 0                  |

<sup>(1)</sup> IQR—Interquartile Range (Q3–Q1); <sup>(2)</sup> MAD—Mean Absolute Deviation.

A detailed analysis of recognition accuracy across six experimental conditions revealed strong interactions between signal origin (AI vs. human) and transmission technology (GSM, VoIP, VoLTE). AI-generated speech was most accurately recognized in the VoIP condition (Mean = 0.71, StdDev = 0.28), whereas human speech was most accurately recognized in GSM (Mean = 0.69, StdDev = 0.28). Notably, human speech transmitted via VoIP dropped below chance level (Mean = 0.42), while AI speech transmitted via VoLTE showed severely impaired recognizability (Mean = 0.27).

These results indicate that different transmission technologies asymmetrically affect the perceptual cues of AI-generated and human speech. GSM appears to preserve human speech cues most effectively, VoIP selectively enhances discriminability of AI-generated signals, and VoLTE disproportionately degrades AI while maintaining moderate recognizability of human speech.

### 5.6. Impact of Participant Characteristics

The study analyzed several demographic and behavioral factors to determine if specific groups are more resilient to voice cloning (see Figure 9).

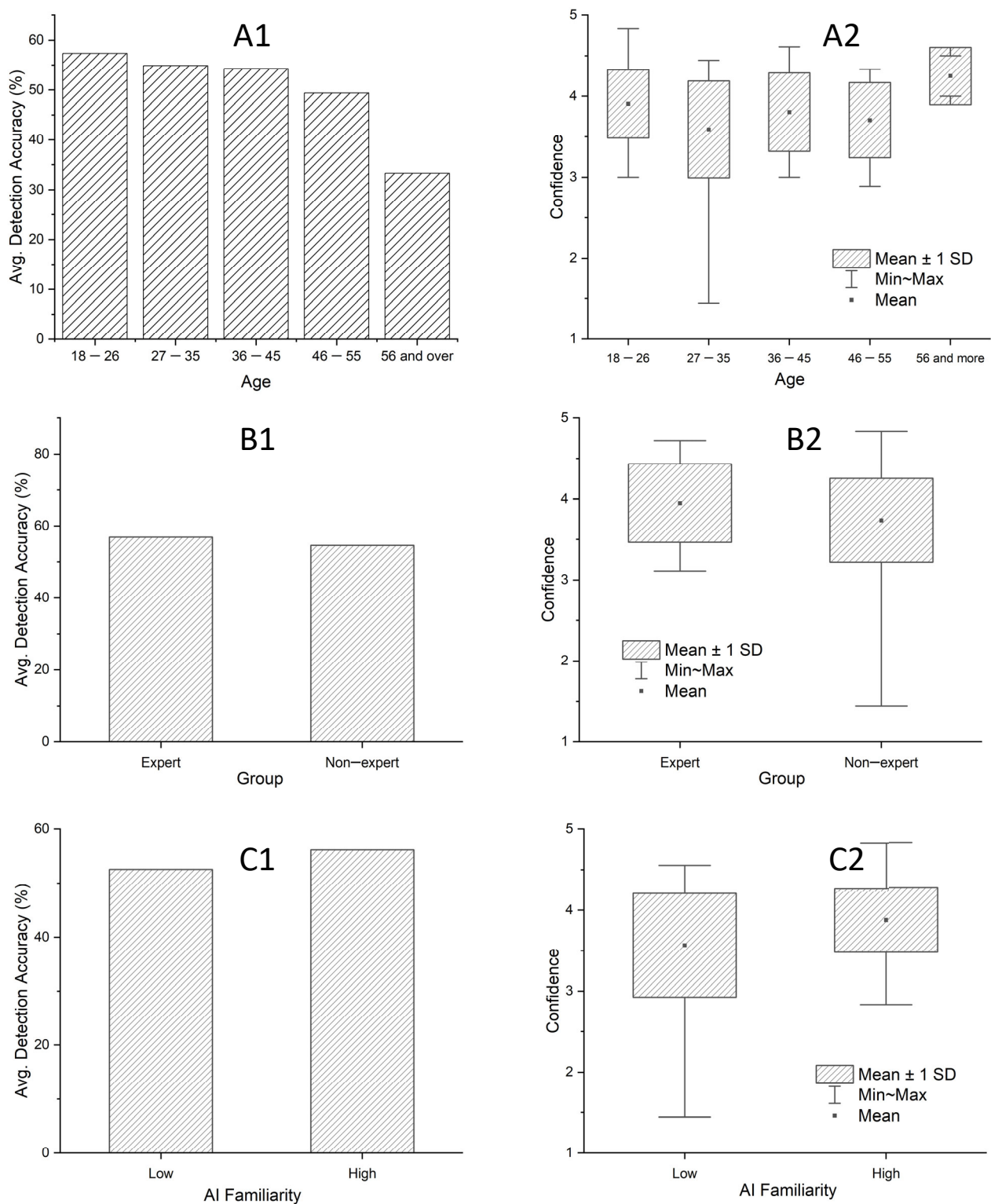


**Figure 9.** Detection accuracy and confidence by participant groups: (A1) Accuracy across age cohorts; (A2) Confidence across age cohorts; (B1) Accuracy for experts vs. non-experts; (B2) Confidence for experts vs. non-experts; (C1) Accuracy by AI familiarity; (C2) Confidence by AI familiarity.

### 5.6.1. Age Effect

Accuracy exhibited a clear negative correlation with age. The youngest cohort (18–26 years) achieved the highest accuracy (56.5%), while participants over 46 years old dropped to 50.3%. Interestingly, confidence levels did not decrease proportionally with age, suggesting that older listeners may be more vulnerable due to a combination of lower detection skills and maintained subjective certainty.

### 5.6.2. Expertise and AI Familiarity

The “Expert” group (audio engineering students) did not show a statistically significant advantage in detection accuracy (55.7%) compared to non-experts (53.9%). However, experts reported significantly higher confidence ratings (4.02 vs. 3.68). Similarly, frequent users of AI tools expressed higher confidence but achieved only a marginal 3% improvement in accuracy. These results point to an “Expertise Overconfidence Effect”, where specialized knowledge increases subjective certainty without necessarily providing immunity to high-quality audio deepfakes.

### 5.6.3. Impact of Telecommunication Channel and Voice Cloning Experience

Table 6 shows the differences in detection accuracy, taking into account the division into two groups of users: frequent and occasional users of AI technology. It is clear that users who use this technology more frequently are generally better at distinguishing AI-generated voice from human voice.

**Table 6.** AI Familiarity vs. Detection Performance.

| AI Usage                 | Detection Performance (%) |           |         |            |          |             |
|--------------------------|---------------------------|-----------|---------|------------|----------|-------------|
|                          | GSM/AI                    | GSM/Human | VoIP/AI | VoIP/Human | VoLTE/AI | VoLTE/Human |
| more than once per week) | 63                        | 70        | 74      | 43         | 22       | 64          |
| once per week or less    | 50                        | 68        | 67      | 41         | 36       | 54          |

On the other hand, as shown in Table 7, users’ previous experience with voice cloning had no significant impact on their ability to distinguish between the two types of voice samples.

**Table 7.** Voice Cloning Experience vs. Detection Performance.

| User’s Knowledge of Voice Cloning | Detection Performance (%) |           |         |            |          |             |
|-----------------------------------|---------------------------|-----------|---------|------------|----------|-------------|
|                                   | GSM/AI                    | GSM/Human | VoIP/AI | VoIP/Human | VoLTE/AI | VoLTE/Human |
| had contact                       | 65                        | 63        | 78      | 20         | 31       | 65          |
| had no contact                    | 57                        | 71        | 70      | 47         | 26       | 60          |

### 5.7. Automatic Baseline Detection Results

The automatic detection results across different feature and training set configurations are summarized in Table 8.

**Table 8.** Automatic baseline detection accuracy across model pipelines and training scenarios.

| Model Pipeline       | Training Data | Accuracy (%) |
|----------------------|---------------|--------------|
| MFCC + Random Forest | Clean         | 61.1         |
| MFCC + Random Forest | Augmented     | 61.1         |
| Wav2Vec2 + SVC       | Clean         | 66.7         |
| Wav2Vec2 + SVC       | Augmented     | 83.3         |

The classical MFCC-RF baseline yielded a constant accuracy of 61.1% across both training configurations. This performance baseline reflects the sensitivity of handcrafted spectral features to low-bitrate compression artifacts, which in this case led to the misclassification of authentic degraded samples as synthetic speech.

Conversely, the self-supervised Wav2Vec2-SVC pipeline demonstrated a higher capacity to handle channel variations when augmented. While the model trained strictly on clean data achieved an accuracy of 66.7%, incorporating channel-simulation data augmentation into the training process increased the evaluation accuracy to 83.3%.

An error analysis of the top-performing configuration (Wav2Vec2 + SVC, Augmented) showed a 100% detection rate for the synthetic samples within the evaluation set, resulting in a 0% False Negative Rate (FNR) across the target GSM, VoIP, and VoLTE transmission conditions. The remaining 16.7% classification error consisted entirely of False Positives, where authentic human speech samples were misclassified as deepfakes. These errors occurred exclusively on files subjected to either GSM or VoIP simulated conditions.

### 5.8. Statistical Analysis

The main goal of the statistical analysis of the obtained results was to answer two key questions:

- Question 1: Do users recognize AI or humans better?
- Question 2: Does the network environment influence decision accuracy?

The answer to the first question is: no significant relationship was found between performance in AI and human recognition tasks.

In order to answer the second question, a more in-depth statistical analysis was carried out. The nonparametric Kruskal–Wallis test was used to examine the impact of the environment (GSM/VoIP/VoLTE) on decision accuracy (the data are not normally distributed). The Kruskal–Wallis test result (Chi-Square = 50.84, DF = 2,  $p < 0.0001$ ) indicates that there are statistically significant differences between the decision outcomes of users evaluating voice samples in three different network environments (the probability that these differences occurred by chance is less than 0.01%). Descriptive statistics are presented in Table 9.

**Table 9.** Kruskal–Wallis test—descriptive statistics.

| Accuracy     | N  | Min | Q1 | Median | Q3 | Max |
|--------------|----|-----|----|--------|----|-----|
| in GSM [%]   | 95 | 17  | 50 | 67     | 83 | 100 |
| in VoIP [%]  | 95 | 17  | 50 | 50     | 67 | 100 |
| in VoLTE [%] | 95 | 0   | 33 | 50     | 50 | 83  |

Descriptive analysis indicates that the highest voice recognition accuracy was achieved in the GSM environment (median 67%), while lower and comparable values were observed for VoIP and VoLTE (median 50%). Additionally, the VoLTE environment was characterized by the lowest minimum values (0%) and a reduced range of results, suggesting greater susceptibility to signal degradation and classification errors. In the study, the GSM network proved to be the most favorable environment for voice recognition accuracy, VoIP moderate, and VoLTE the weakest and most unstable. The MIN/MAX values represent the extreme results of individual participant observations. A MIN value of 0 means that someone (or a sample of decisions) had a complete misclassification, meaning all their responses were incorrect for the given case (here, the network environment). Similarly, the MAX value indicates the best accuracy of decisions made in a given network environment. Post hoc analysis showed detailed differences between individual result groups corresponding to

different network environments. Details of this analysis (Dunn's Test), presented as paired comparison plot, are shown in Figure 10.

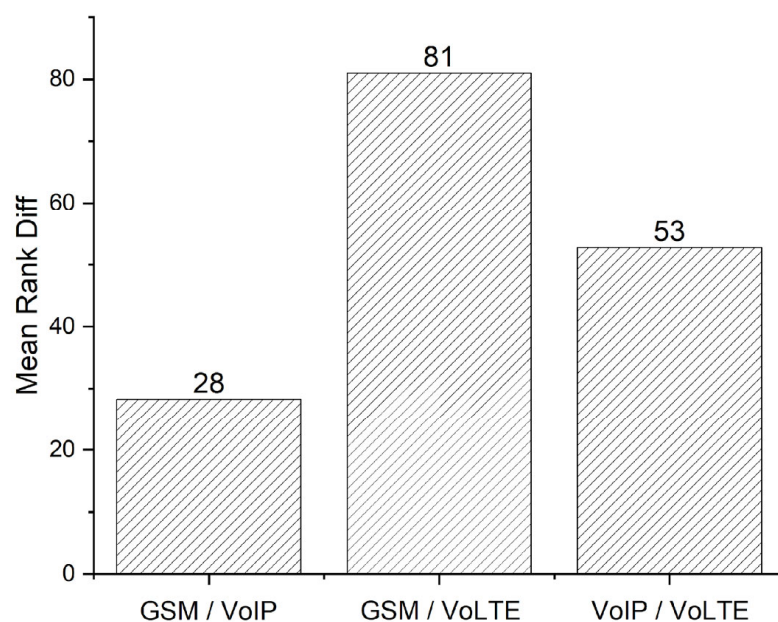


**Figure 10.** Pairwise post hoc comparisons across transmission conditions based on Dunn's test.

Pairwise comparisons using Dunn's test indicated that the magnitude of differences varied across conditions, with the strongest effect observed between GSM and VoLTE, followed by VoIP and VoLTE, and the weakest between GSM and VoIP.

### 5.9. Summary of Key Findings

The empirical evidence collected in this study leads to three primary conclusions:

1. Human detection is unreliable: Mean accuracy (54.8%) is too low to provide any meaningful defense against voice-cloning attacks.
2. High-fidelity channels may increase susceptibility to misclassification: The "Quality-Authenticity Bias" in VoLTE networks actively misleads listeners, pushing accuracy below chance level.
3. Confidence is misleading: There is no reliable correlation between how sure a listener feels and whether they are correct, a gap that is particularly dangerous in "expert" groups.

## 6. Discussion

The present study reveals several critical insights into how humans perceive and misperceive voice-cloned speech transmitted through real-world telecommunication channels. Beyond simple accuracy metrics, the findings highlight a complex interplay between signal degradations, cognitive heuristics, and perceptual expectations. Transmission channels fundamentally alter both the clarity of the cues available to the listener and the subjective criteria used to evaluate voice authenticity.

### 6.1. Analysis of High Detection Performance in the GSM Channel

The highest detection accuracy was observed in the GSM condition (63.7%). This result invites two complementary interpretations, and the reality likely resides at their intersection:

1. Isolation of Prosodic Irregularities: The narrow bandwidth (300–3400 Hz) and AMR-NB compression of the GSM channel eliminate high-frequency components that

typically smooth synthetic speech. By removing these frequencies, the channel may inadvertently expose fundamental prosodic flaws—such as timing rigidity and pitch contour oversmoothing—which remain major limitations of contemporary generative models [9,11]. In this scenario, channel degradation acts as a filter that unmasks synthesized flaws.

2. **Skepticism Toward Degraded Signals:** Conversely, the acoustic degradation inherent to narrowband telecommunication matches pre-existing cognitive expectations regarding artificial speech. Lower bitrates and spectral flattening may trigger a general skepticism, making listeners more inclined to classify any degraded signal as synthetic. In this scenario, the elevated accuracy rate is partially a byproduct of a lower subjective threshold for labeling samples as AI-generated, rather than superior acoustic discrimination.

### 6.2. Perceptual Effects of High-Fidelity Channels (VoLTE)

The drop in accuracy to 44.0% in the VoLTE condition—significantly below chance—points to a systematic failure in human judgment. The wideband AMR-WB codec (50–7000 Hz) provides acoustic clarity that listeners instinctively associate with more natural, live communication. This creates a quality-driven bias, where high acoustic fidelity may lead listeners to overlook subtle synthesis artifacts. This interpretation is further supported by the SDT analysis. VoLTE showed the lowest sensitivity ( $d' = -0.20$ ) and the strongest positive decision criterion ( $c = 0.45$ ), indicating not only weak discriminability but also a systematic tendency to classify samples as human. This pattern supports the interpretation that wideband, high-fidelity transmission can bias listeners toward authenticity judgments even when perceptual evidence is insufficient.

The absence of traditional telephony noise reinforces the listener's prior belief that a clear, professional-sounding voice must be genuine. This finding suggests that as telecommunication infrastructure adopts higher bandwidths (e.g., 5G EVS codecs), increased audio quality will continue to diminish the detectability of voice clones.

### 6.3. Ambiguity in Packet-Based Networks (VoIP)

VoIP yielded intermediate results (56.7%). The characteristic impairments of packet-based networks—such as jitter, packet loss, and PLC (Packet Loss Concealment) artifacts—introduce sporadic discontinuities. These network-induced distortions are perceptually similar to the “glitches” or phase incoherence occasionally produced by neural audio codecs [16].

This creates a state of perceptual interference, where listeners struggle to attribute a perceived artifact to the source (the AI model) or the channel (the network). Consequently, VoIP introduces decision instability, where the listener's criterion shifts from trial to trial, leading to inconsistent performance and lower overall confidence compared to the more “stable” (even if incorrect) environments of GSM and VoLTE. Consistently, the SDT analysis showed a negative decision criterion for VoIP ( $c = -0.43$ ), indicating a bias toward “AI” responses, which is compatible with the interpretation that packet-based distortions may be mistaken for synthetic artifacts.

### 6.4. Comparison with Existing Literature

The overall accuracy of 54.8% aligns with recent findings by Barrington et al. [9] and Mai et al. [10], who noted that human detection is rapidly approaching chance level as models like VALL-E or NaturalSpeech improve. However, our results provide a necessary correction to studies conducted solely with studio-quality audio. While Mai et al. [10] reported accuracies near 70% in clean conditions, our study demonstrates that real-world telecommunication channels act as a leveling field, either by masking synthetic cues (VoLTE)

or by inducing biases that inflate or deflate accuracy based on signal quality rather than true discrimination.

#### 6.5. Comparison with Automatic Detection Baseline

The comparative analysis highlights opposing trends between human participants and the automated baseline. While human deepfake detection accuracy was lowest in wideband channels like VoLTE, the channel-augmented Wav2Vec2 model remained robust under these conditions. This indicates that self-supervised neural representations can capture voice-cloning artifacts within high-quality transmission channels that typically mislead human listeners. Consequently, the automated baseline performs reliably precisely where synthetic speech closely matches the perceptual quality of a natural voice.

Importantly, the results demonstrate that even an unrefined, out-of-the-box model architecture—trained on a restricted dataset without extensive fine-tuning—performs comparably to or significantly outperforms human judges across identical audio stimuli. While the human auditory system remains highly vulnerable to specific channel configurations, deep contextual neural embeddings provide a more consistent detection baseline under changing transmission constraints.

Finally, the relatively small size of the evaluated audio dataset was an intentional requirement of the experimental design. Conducting a rigorous human listening test imposes strict limits on the total duration of the audio stimuli to prevent auditor fatigue and ensure consistent subjective judgment. Despite the limited number of samples, this database serves its direct purpose: establishing a strict, side-by-side comparison between human perception and automated systems under identical channel impairments.

#### 6.6. Implications for Security and Trust

The mismatch between high confidence and low accuracy, particularly in high-quality channels, presents a severe challenge for cybersecurity:

- **The Overconfidence Risk:** Participants with technical backgrounds expressed higher confidence but did not significantly outperform the general population. This miscalibration between subjective confidence and actual performance represents a primary vulnerability to targeted social engineering attacks (vishing).
- **Infrastructure Vulnerability:** The fact that improved signal quality (VoLTE) leads to worse detection suggests that modernizing networks may unintentionally facilitate deepfake-enabled fraud.

#### 6.7. Core Interpretation: The Heuristic of Sound Quality

The findings of this study indicate that human discrimination of synthetic speech is primarily mediated by perceived audio quality rather than the direct detection of generative features. This evaluation framework manifests across three distinct channel conditions:

1. Low fidelity (GSM) triggers a categorical bias toward “Artificial,” resulting in higher (but potentially lucky) accuracy.
2. High fidelity (VoLTE) triggers a categorical bias toward “Natural,” resulting in systematic deception.
3. Ambiguity (VoIP) creates a confusion of sources, where channel noise masks synthesis errors.

As generative models continue to close the prosodic gap, and telecommunication networks continue to close the quality gap, human perceptual judgment alone may become increasingly unreliable as a primary detection mechanism. This necessitates a transition toward multi-layered, automated verification systems that do not rely on human perceptual judgment.

The findings should be interpreted in light of the controlled dataset design, which includes a limited number of Polish speakers, a narrow demographic range, a single commercial voice cloning system, and a restricted number of stimuli selected to maintain feasible listening-test duration.

## 7. Conclusions and Future Work

This study demonstrates that human listeners are inherently ill-equipped to detect modern voice-cloned speech when it is transmitted through realistic telecommunication channels. Based on the analysis of 1710 judgments from 95 participants, the overall detection accuracy of 54.8% confirms that human discriminatory ability has been effectively eroded by the rapid advancement of generative speech models. The supplementary SDT analysis further showed that these effects are not limited to reduced discriminability alone, but also involve channel-dependent shifts in response bias, particularly toward “Human” responses in VoLTE and toward “AI” responses in VoIP.

The experimental results yield three primary conclusions. First, channel-dependent acoustic fidelity significantly modulates human judgment. Narrowband channels (GSM), despite their technical limitations, can aid detection (63.7%) by exposing prosodic rigidity, whereas wideband channels (VoLTE) degrade performance below chance level (44.0%) due to a quality-driven bias where listeners equate high acoustic clarity with authenticity. Second, a severe discrepancy exists between listener confidence and accuracy; participants exhibited the highest subjective certainty in wideband scenarios where they were most frequently deceived. Third, human judgment alone may be insufficient as a primary security control against voice-cloning attacks, as technical expertise or prior familiarity with synthetic media provided no significant advantages against deepfake-enabled impersonation.

Consequently, future security frameworks must transition from human perception toward automated, multi-layered algorithmic verification integrated into the telecommunication infrastructure. This study addressed this need by implementing and validating automatic baseline pipelines alongside the human perceptual experiment. Future work will expand this framework by training specialized anomaly detection architectures to identify reconstruction errors in synthetic speech. Additionally, subsequent studies will evaluate model resilience against emerging 5G codecs and advanced adaptive networks. Finally, while the current study utilized a simplified VoIP impairment model intentionally designed to isolate specific perceptual effects, future research will expand this scope by incorporating standardized network models and evaluation frameworks, specifically integrating the E-model, POLQA metrics, and the Opus codec to validate both human and machine performance under industry-standard telecom conditions.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/acoustics8020041/s1>.

**Author Contributions:** Conceptualization, J.W. and M.L.; methodology, J.W., M.L. and J.K.; software, J.W.; validation, J.W., M.L. and J.K.; formal analysis, J.W. and J.K.; investigation, J.W.; resources, J.W. and M.L.; data curation, J.W.; writing—original draft preparation, M.L.; writing—review and editing, J.W. and J.K.; visualization, J.K.; supervision, M.L. and J.K.; project administration, J.W. and M.L.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study was conducted in accordance with the principles of the Declaration of Helsinki. According to institutional and national guidelines, this type of non-invasive study based on anonymous perceptual data does not require formal ethical approval.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data presented in this study are available as a Supplementary File to this article.

**Acknowledgments:** The authors would like to acknowledge the participants who took part in the listening experiments for their time and engagement. The authors also thank colleagues and student volunteers for their assistance with participant recruitment and for valuable technical discussions during the experimental design phase. During the preparation of this manuscript, the authors used Generative AI tools (Microsoft 365 Copilot, based on GPT-5 chat model) to support text editing, language refinement, and structural organization of the manuscript. The tool was not used for data generation, data analysis, experimental design, or interpretation of results. All outputs produced with the assistance of generative AI were carefully reviewed, edited, and validated by the authors, who take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; et al. NaturalSpeech: End-to-end text to speech synthesis with human-level quality. *arXiv* **2022**, arXiv:2205.04421. [\[CrossRef\]](#)
2. Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, Y.; Leng, Y.; Song, K.; Tang, S.; et al. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv* **2024**, arXiv:2403.03100. [\[CrossRef\]](#)
3. Jia, Y.; Zhang, Y.; Weiss, R.J.; Wang, Q.; Shen, J.; Ren, F.; Chen, Z.; Nguyen, P.; Pang, R.; Moreno, I.L.; et al. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 4485–4495. [\[CrossRef\]](#)
4. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv* **2023**, arXiv:2301.02111. [\[CrossRef\]](#)
5. Barrington, S.; Cooper, E.A.; Farid, H. People are poorly equipped to detect AI-powered voice clones. *Sci. Rep.* **2025**, *15*, 11004. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Mai, K.T.; Bray, S.; Davies, T.; Griffin, L.D. Warning: Humans cannot reliably detect speech deepfakes. *PLoS ONE* **2023**, *18*, e0285451. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Milewski, K.; Zaporowski, S.; Czyżewski, A. Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice. *Electronics* **2023**, *12*, 4458. [\[CrossRef\]](#)
8. ITU-T. *Recommendation P.800: Methods for Subjective Determination of Transmission Quality*; International Telecommunication Union: Geneva, Switzerland, 1996.
9. Hunt, A.J.; Black, A.W. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, GA, USA, 7–10 May 1996; Volume 1, pp. 373–376. [\[CrossRef\]](#)
10. Zen, H.; Tokuda, K.; Black, A.W. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *51*, 1039–1064. [\[CrossRef\]](#)
11. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499. [\[CrossRef\]](#)
12. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783. [\[CrossRef\]](#)
13. Kim, J.; Kong, J.; Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv* **2021**, arXiv:2106.06103. [\[CrossRef\]](#)
14. Yamagishi, J.; Wang, X.; Todisco, M.; Sahidullah, M.; Patino, J.; Nautsch, A.; Liu, X.; Lee, K.A.; Kinnunen, T.; Evans, N.; et al. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *arXiv* **2021**. [\[CrossRef\]](#)
15. 3GPP. *Mandatory Speech CODEC Speech Processing Functions; AMR Speech CODEC; General Description; Technical Specification (TS) 26.071*; 3rd Generation Partnership Project (3GPP): Valbonne, France, 2022.
16. Raake, A. *Speech Quality of VoIP: Assessment and Prediction*; John Wiley & Sons: Hoboken, NJ, USA, 2007.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.