Human Perception of Arabic Fake Audio

Elsayed Issa
Purdue University
West Lafayette, IN, USA
esissa@purdue.edu

Yi-Fang Cheng
Purdue University
West Lafayette, IN, USA
cheng627@purdue.edu

Abstract

Audio deepfake research has focused on a few high resource languages, leaving human perception of synthesized Arabic speech unexplored. We present the current study of how human listeners perceive Arabic audio deepfakes. Thirty eight native speakers judged Modern Standard Arabic stimuli in two formats generated with proprietary and open-source models. The two formats are a binary task pairing genuine clips with text-to-speech counterparts and a unary task requiring single-clip authenticity judgments of voice-cloned political speech. The results showed that listeners performed only moderately above chance in both formats (binary: 62.9%; unary balanced accuracy: 64.6%). In the unary task, listeners showed a systematic bias toward judging audio as authentic ($d' = 0.77$, $c = 0.27$), accepting 45% of synthetic clips as real. Confidence was positively related to accuracy in the binary condition but inversely related in the unary condition. Because the two conditions also differ in synthesis method, speakers, and stimuli, we treat this as a descriptive contrast between conditions rather than an effect of task format. An exploratory coded analysis of participants' rationales suggested that reasoning from subjective certainty was associated with lower accuracy and concrete perceptual cues with higher accuracy, while no cue category reliably predicted confidence. These findings suggest that individual listeners are an unreliable safeguard against Arabic fake audio detection, and motivate future direct comparisons with automatic detectors.

1 Introduction

Recent advances in generative artificial intelligence have produced synthetic speech that is increasingly indistinguishable from human speech (Lavan et al., 2025). State-of-the-art (SOTA) voice synthesis models (Li et al., 2025) have lowered the technical barrier to such an extent that virtually anyone can generate cloned or fully synthetic voices that listeners cannot reliably distinguish from authentic recordings. When voices that faithfully reproduce human timbre, prosody, and conversational coherence circulate alongside manipulated images and video, they become potent vectors for mis- and disinformation (Shoaib et al., 2023). The downstream consequences are substantial: synthetic media can damage the reputations of public figures (Alanazi et al., 2025), misrepresent political leaders for partisan ends (Metz, 2019; Weiner and Lawrance, 2023), destabilize civic discourse, and propagate fabricated news narratives.

In response, the research community has launched a range of initiatives to combat audio deepfakes, including the ASVspoof (Automatic Speaker Verification Spoofing and Countermeasures) challenge (Wu et al., 2015; Kinnunen et al., 2017; Liu et al., 2023), the ADD (Audio Deep Synthesis Detection) challenge (Yi et al., 2022, 2023), and the VoicePrivacy Challenge (Tomashenko et al., 2020), among others. However, these efforts have focused almost exclusively on a narrow set of high-resource languages such as English, Mandarin, and Spanish, leaving Arabic comparatively underexamined. A small but growing body of work has begun to address this gap: Issa (2026) and Almutairi and Elgibreen (2023) studied detection in Modern Standard Arabic (MSA); Maged et al. (2025) introduced a multi-dialect spoof-speech benchmark; Awaad et al. (2024) examined real-time detection for Egyptian Arabic; and Kuckreja et al. (2025) released ArEnAV, a large-scale Arabic–English audio-visual deepfake dataset.

Despite this progress, existing work on Arabic audio deepfakes has been confined to machine-based detection, while human perception of synthesized Arabic speech remains unexplored. This omission is consequential: in real-world settings, humans are often the last line of defense against deceptive audio, and understanding where listeners

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 35–49
November 10, 2026 ©2026 Association for Computational Linguistics

succeed or fail is essential for designing effective detection systems, public-awareness campaigns, and policy interventions. To the best of our knowledge, this paper presents one of the few studies of human perception of Arabic audio deepfakes.

2 Related work

2.1 Audio Deepfakes

Audio deepfakes encompass both the synthesis of artificial voices and the manipulation of authentic speech (Khanjani et al., 2023). Synthetic speech is generated primarily through two pathways: text-to-speech (TTS) and voice conversion (VC) (Li et al., 2025). TTS systems produce natural-sounding speech directly from text, whereas VC systems transform a source speaker's utterance to match the timbre and style of a target speaker while preserving its linguistic content (Arik et al., 2018). A closely related technique, voice cloning, infers a speaker's vocal identity from only a few samples and synthesizes novel utterances in that voice (Arik et al., 2018). Beyond algorithmic synthesis, adversarial attacks can deliver spoofed audio to a device microphone through physical transduction paths. Such attacks target machine listeners rather than humans, aiming to mislead automatic speech recognition (ASR) systems, voice assistants, or speaker-verification models by exploiting model and sensor vulnerabilities (Liu et al., 2023; Yamagishi et al., 2021). A canonical example is the replay attack, in which a legitimate recording is captured and re-emitted to impersonate a user and defeat biometric voice-authentication systems (Li et al., 2017).

2.2 Arabic Audio Deepfake Detection

Arabic anti-spoofing research remains limited. Almutairi and Elgibreen (2023) proposed an Arabic deepfake detector based on self-supervised representations (Wav2Vec 2.0), trained on a small, single-speaker MSA synthetic corpus they constructed, while Awaad et al. (2024) studied real-time detection for Egyptian Arabic, examining short-duration clips in a streaming, low-latency setting. Complementing these efforts, Kuckreja et al. (2025) released ArEnAV, a large-scale Arabic–English audio–visual deepfake dataset with intra-utterance code-switching, dialectal variation, and monolingual Arabic, intended to stress multimodal detectors beyond monolingual domains. Finally, Maged et al. (2025) introduced a multi-dialect spoof-speech benchmark, pairing bona fide audio from the Casablanca corpus (Talafha et al., 2024) with synthetic spoofs for eight dialects (Algerian, Egyptian, Emirati, Jordanian, Mauritanian, Moroccan, Palestinian, and Yemeni) generated with open-source TTS and VC systems. Issa (2026) introduced AFAD-MSA, a curated corpus of authentic and synthetic Modern Standard Arabic speech in which the synthetic subset is generated with four proprietary TTS and VC systems, and established reference detectors by training AASIST models against Wav2Vec 2.0 and Whisper baselines. The best detector achieved near-perfect in-domain accuracy but degraded under cross-dataset evaluation.

2.3 Human Perception

Fewer studies have examined how well humans detect speech deepfakes. Watson et al. (2021) asked 53 English-speaking college students to judge eight clips from a MelGAN-based TTS model, finding that shorter and more grammatically complex clips were identified more accurately and that familiarity with the concept of deepfakes did not help. Müller et al. (2022) had 410 participants compete against an AI detector on ASVspoof 2019 samples (Wang et al., 2020) and found that humans and the detector performed similarly. Mai et al. (2023) tested 529 English and Mandarin listeners in two formats: a unary task, judging single clips as real or fake, and a binary task, identifying the fake within a pair. Accuracy was about 73% in the unary task and 86% in the binary task, familiarization with example deepfakes raised accuracy by less than four percent, and, contrary to Watson et al. (2021), shorter clips were not easier to detect. For Arabic, Kuckreja et al. (2025) included a user study with ArEnAV in which participants identified real and fake videos with around 60% accuracy; to our knowledge, no study has examined audio-only Arabic deepfakes.

Together, these studies suggest that humans are unreliable deepfake detectors and that training individuals yields only marginal gains. More promising are collective and machine-assisted approaches: Groh et al. (2022) showed that crowd aggregation and access to model predictions can outperform humans or models alone, though erroneous model outputs often misled participants, and Mai et al. (2023) found that aggregated human judgments rival top automated detectors on speech while being more robust out of domain.

3 Methods

3.1 Participants

The study initially involved 42 Arabic native speakers aged 20 to 40. The present analysis used the complete-response subset of the data. As 4 participants did not complete the full experiment, the corresponding data were excluded from the present analysis. This resulted in a total of 38 participants in the final analysis.

3.2 Audio Materials and Stimuli

We compiled our stimulus set from two complementary sources. For the text-to-speech (TTS) condition, we drew nine genuine recordings from the CommonVoice dataset (Ardila et al., 2020) and the Arabic Speech Corpus (Halabi et al., 2016). Each recording served as the reference for synthesizing a fake counterpart, with three recordings assigned to each of three SOTA proprietary TTS models: ElevenLabs, ResembleAI, and MiniMax (Table 8). For the voice cloning (VC) condition, we extracted three reference clips (7–10 s) from publicly available speeches delivered by three prominent Arab political figures, one clip per figure (Table 7). Each reference clip was used both as a genuine stimulus and as the speaker prompt for cloning.

Cloning was zero-shot. Each of four synthesis systems (two proprietary and two open-source) received the reference clip together with its transcript as the text to be synthesized, yielding twelve cloned variants. No source speaker or voice conversion was involved. Each cloned variant reproduced the transcript of its genuine reference clip, so real and synthetic clips of the same figure were identical in content: the figures say only what they actually said in the original address, and the content itself is neutral and plausible. The unary task presented these clips one at a time, and participants were asked to judge each as genuine or fake.

For the TTS condition, genuine clips ranged from 3 to 7 s (Table 8), and each synthetic counterpart was generated from the same transcript. All stimuli were therefore short (3–10 s). Table 1 summarizes the stimulus materials and Appendix A describes the generation details. It is worth noting that the unary task is imbalanced by design (12 synthetic vs. 3 genuine clips), so a listener who always responded "fake" would reach 80% raw accuracy. We therefore report balanced accuracy and signal detection measures for this task alongside raw accuracy.

Table 1: Summary of stimulus materials. The TTS condition supports the binary pairwise task; the VC condition supports the unary authenticity task.

Condition	Genuine	Synthetic	Models
TTS	9 clips	9 fake counterparts	ELEVENLABS, MINIMAX, RESEMBLEAI
VC	3 clips	12 cloned variants	FISHAUDIO, HUME, MINIMAX, XTTS-V2

Two design choices warrant explanation. First, the political figures were selected for their broad recognizability among Arabic speakers and, critically, their consistent use of MSA, the variety under investigation, in official addresses. Using recognizable figures mirrors the most consequential real-world use of voice cloning, but it raises the possibility that listeners draw on familiarity with the speaker. Two features of the design limit this concern: the binary TTS task uses anonymous speakers, providing a condition in which familiarity cannot contribute, and because cloned and genuine clips share the same transcript, knowledge of what a figure has said cannot by itself reveal authenticity. Any benefit of familiarity is thus restricted to knowledge of the voice itself.

In addition, each unary trial displayed a photograph of the speaker (see Figure 6). Because the cloned voices belonged to public figures, the images were intended as an additional cue. Identifying the speaker allows listeners to compare the clip against their knowledge of that person's voice, which gives them the best opportunity to detect a clone. The images were identical for genuine and cloned clips of the same figure and therefore carry no information about authenticity in themselves. Since no comparable identity information exists for the anonymous TTS speakers, images were shown only in the unary condition, and their effect cannot be separated from the other differences between conditions (see Section 5). Second, the set of synthesis models differs between the TTS and VC conditions because several proprietary systems (e.g., ElevenLabs) impose policy restrictions on voice cloning. We deliberately favored proprietary models where possible in order to maximize the perceptual quality of the synthetic stimuli, thereby ensuring that human listeners would have to engage in genuine effortful discrimination rather than detect obvious artifacts.

3.3 Data Collection Procedure

The experiment was conducted online via a Qualtrics link through which participants received instructions and completed two tasks (see Figure 6 for an example from the survey). The binary task consisted of the first 9 questions, where listeners selected which of an A/B pair was fake. The unary task consisted of the remaining 15 questions, where listeners judged a single audio as real or fake. The order of the two tasks and the order of the questions within each task were randomized. Each question was followed by a confidence-rating item and a free-text response prompt in which participants explained their reasoning behind their response. Participants were allowed to listen to the clips as often as they wished and did not receive feedback indicating whether their responses were correct.

3.4 Statistical Analysis

The dataset contained 912 observations: 342 responses from the binary task (38 participants × 9 questions) and 570 responses from the unary task (38 participants × 15 questions). Accuracy was coded as 1 for a correct response and 0 otherwise. Confidence was measured using a 10-point Likert scale and was z standardized before fitting the regression models.

We first summarized accuracy descriptively by task format, confidence rating, and model condition. We then fitted mixed-effects logistic regression models to account for the repeated-measures structure of the data. The primary model predicted response accuracy from task format, standardized confidence, and their interaction, with random intercepts for participants and questions, which were the maximal random effects that allowed the algorithm to find a stable solution (Bates et al., 2015). To examine model-level variation, we fitted additional mixed-effects logistic regression models separately for the binary and unary tasks. For the binary task, accuracy was modeled as a function of the TTS model used to generate the fake counterpart. For the unary task, accuracy on generated items was modeled as a function of the VC model.

We additionally conducted signal detection theory (SDT) analysis (Hautus et al., 2021) to distinguish listeners' ability to discriminate real from fake audio from their response bias. Treating fake audio as the signal, we computed the hit rate, false-alarm rate, sensitivity d', and criterion c. This allowed us to examine whether errors reflected limited discrimination, a systematic tendency to respond "real" or "fake," or both. We also computed a confidence-weighted receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUROC) (Swets, 2014). All statistical analyses were conducted in R (R Core Team, 2024).

3.5 Participants' Reasoning

Participant responses were in either Arabic or English. Arabic responses were translated into English by an Arabic native speaker and verified by the first author. An annotation scheme was created based on the cues listeners report (e.g., naturalness, prosody, and linguistic anomalies) (see Table 9 in the Appendix C). The scheme comprised nine categories and each response was coded multi-label, as responses frequently cited more than one cue. Coding was done manually by two annotators, both Arabic native speakers, who independently coded all responses. Agreement was substantial overall (Landis and Koch, 1977) (mean Cohen's $\kappa = 0.76$), ranging from 0.52 for audio quality to 0.89 for naturalness (see Table 9 for per-category values). Disagreements were resolved through discussion to produce the final labels used in the analyses.

To examine which cues were associated with detection performance, we entered each of the nine cue categories as a predictor in mixed-effects logistic regression models of response accuracy, with random intercepts for participants and questions. Cues were modeled both individually and in a single combined model, with each category coded as a binary indicator of whether it appeared in a given response. We additionally fitted linear mixed-effects models predicting confidence ratings from each cue category, using the same random-effects structure. Because coding was multi-label, "subjective certainty" denotes any response in which the participant appealed to an impression or feeling of certainty (e.g., "it feels real", "I'm sure"). This code could co-occur with other cues. Given the number of cue categories tested, the cue analyses are exploratory. We report uncorrected p-values and note which effects survive Holm correction across the nine univariate models.

4 Results

4.1 Overall Performance

Participants performed moderately above chance in both task formats. Binary forced-choice accuracy was 62.9% (chance = 50%). Raw unary accuracy

was 58.6% (Figure 1), but this figure is dominated by the 12 synthetic items (80% of unary trials) and falls below the 80% that a constant "fake" response would yield. Averaging accuracy over the two classes gives a unary balanced accuracy of 64.6% (Table 2), comparable to binary accuracy. The mixed-effects model likewise found no difference between conditions at average confidence ($\beta = -0.177$, $p = .748$; Table 3). We therefore do not interpret either format as yielding higher overall accuracy.

	Binary	Unary
Raw accuracy	62.9%	58.6%
Accuracy on fake items	–	54.6%
Accuracy on real items	–	74.6%
Balanced accuracy	62.9%[a]	64.6%
Constant "fake" (raw / bal.)	50% / 50%	80% / 50%

Table 2: Accuracy by condition. See Figure 5.

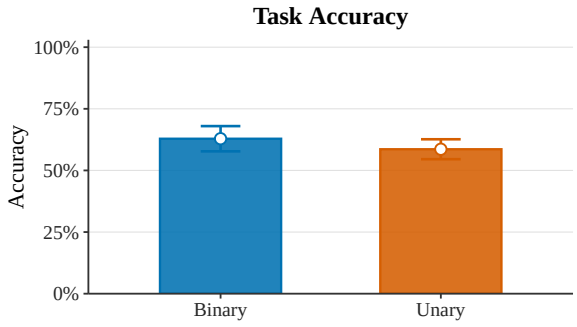


Figure 1: Aggregate binary and unary task accuracy, with capped 95% confidence intervals. Unary values are raw accuracy; see Table 2 for balanced accuracy.

Figure 5 in Appendix D provides confusion matrices to clarify the error structure. In the binary task, participants selected the fake clip correctly in approximately 63% of trials. In the unary task, participants were more accurate for real than fake clips. 75% of real clips were correctly classified as real, whereas 55% of fake clips were correctly classified as fake. Thus, 45% of fake clips were judged as real. This indicates that many synthetic samples were perceived as authentic in the single-clip judgment setting.

4.2 Confidence Rating

To examine whether participants' confidence ratings were associated with their decisions, we compared confidence patterns across task formats and response types. Mean confidence was slightly higher in the unary task (7.66) than in the binary task (7.33), despite lower raw accuracy in the unary task (Figure 2). The pattern suggests that higher confidence did not necessarily correspond to higher accuracy. Separating confidence by fake/real type provides a more diagnostic view. Binary-fake responses are correct fake selections, binary-true responses are incorrect true selections, unary-fake items are generated items, and unary-true items are genuine items.

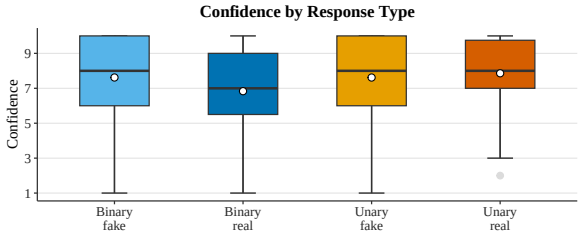


Figure 2: Accuracy and confidence patterns. Confidence distributions separated by task and fake/real type. In the binary task, binary-fake corresponds to correct selections of the fake clip, whereas binary-true corresponds to incorrect selections of the true clip.

Considering whether confidence level predicted accuracy, we fitted a mixed-effects logistic regression model with task format, standardized confidence, and their interaction as fixed effects (Table 3). The model showed a significant effect of confidence ($\beta = 0.418$, $p = .001$). The interaction between task format and confidence was also significant ($\beta = -0.777$, $p < .001$), indicating that the relationship between confidence and accuracy differed by task format. Specifically, confidence was positively related to accuracy in the binary task, meaning that higher confidence ratings were associated with higher accuracy. However, this relationship was reversed in the unary task, where higher confidence was associated with lower accuracy.

These results show that the confidence–accuracy relationship differed between the two experimental conditions. This difference cannot be attributed to task format or to the availability of a reference clip alone, since the conditions also differ in synthesis method (TTS vs. voice cloning), synthesis systems, speaker familiarity, speaker images, recording sources, and clip duration (see Section 5). Nor does it reflect the class imbalance of the unary task: the inverse relationship holds when the unary model is fit on synthetic items only (Table 5, $\beta = -0.485$, $p = .001$). In the unary condition, listeners were thus often confident when accepting cloned clips as genuine.

Table 3: Logistic mixed-effects model predicting response accuracy from task format and confidence.

	β [95% CI]	SE	z	p
Intercept (Task Binary)	0.704 [–0.154, 1.563]	0.438	1.608	.108
Task Unary	–0.177 [–1.257, 0.904]	0.551	–0.321	.748
Confidence (z-scaled)	0.418 [0.168, 0.669]	0.128	3.280	**.001**
Task Unary × Confidence (z-scaled)	-0.777 [–1.100, –0.454]	0.165	–4.713	**<.001**

4.3 Model Condition

Because each TTS system was applied to a different set of three genuine clips (Table 8), and the ResembleAI items were drawn from a different corpus, system identity is confounded with item in the binary task. The model-level comparisons below are therefore descriptive and should not be read as a ranking of the systems. Considering model condition, accuracy varied descriptively across synthesis models, as shown in Figure 3. In the binary task, ElevenLabs yielded the highest accuracy (71.1%), whereas ResembleAI yielded the lowest accuracy (48.2%). However, in the mixed-effects model with random intercepts for participants and questions, neither MiniMax nor ResembleAI differed significantly from the reference model (ElevenLabs) (MiniMax: $\beta = -0.055$, $p = .949$; ResembleAI: $\beta = -1.390$, $p = .106$), although ResembleAI showed significantly lower accuracy in a simpler model with only participant-level random intercepts ($\beta = -1.196$, $p < .001$). This suggests that the apparent model-level variation in the binary task largely reflects item difficulty. Table 4 presents the logistic mixed-effects model for the binary task with participant- and question-level random intercepts.

In the unary task, open-source models were detected more accurately, 58.3%, than proprietary models, 50.9%, yet this difference was not statistically significant. ($\beta = -0.797$, $p = .423$). However, the identity of the model contributed to the performance of the participants. Compared to FishAudio, XTTS was significantly more likely to be detected as fake ($\beta = 3.152$, $p = .001$). Hume ($\beta = 1.497$, $p = .107$) and MiniMax ($\beta = 0.129$, $p = .890$) did not differ significantly from FishAudio. Table 5 presents the logistic mixed-effects model for the unary task. Full Tukey-adjusted pairwise contrasts are reported in Appendix D in Tables 10 and 11.

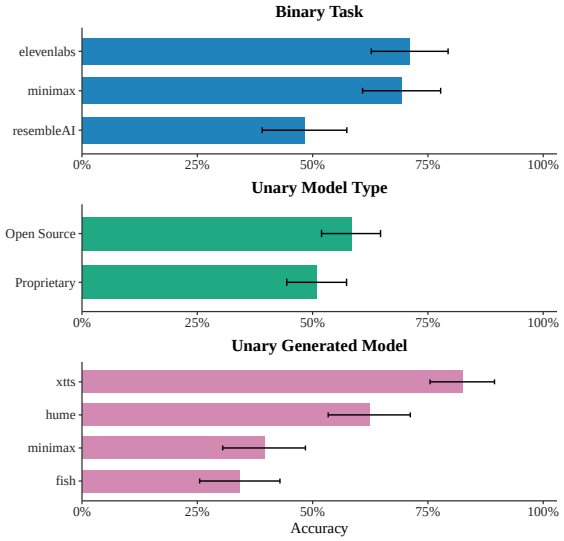


Figure 3: Accuracy by model and model type.

4.4 Unary Sensitivity

As a substantial proportion of fake items were judged as real, we conducted SDT and ROC analyses for the unary task. We treated fake audio as the signal. The hit rate was 54.6% and the false-alarm rate was 25.4%. This yielded $d' = 0.77$, indicating above-chance but moderate discrimination between real and fake audio. The criterion estimate was $c = 0.27$, indicating a conservative bias against saying "fake." This helps explain why unary accuracy was modest: participants showed some discrimination ability, but they also had a response tendency that increased misses for fake items (i.e., judging fake items as real). Unlike raw accuracy, d' and c are unaffected by the unequal numbers of real and fake items, although the false-alarm rate rests on only three genuine clips and is correspondingly less precise.

The ROC analysis showed a similar pattern. Each unary response was converted into a fake-detection score. "Fake" responses received positive confidence values from +1 to +10, and "real" responses received negative values from –1 to –10.

Table 4: Logistic mixed-effects model predicting correctness by model for binary task.

	β [95% CI]	SE	z	p
Intercept (Model ElevenLabs)	1.277 [0.029, 2.542]	0.636	2.006	.045
Model MiniMax	–0.055 [–1.751, 1.641]	0.865	–0.063	.949
Model ResembleAI	–1.390 [–3.077, 0.297]	0.861	–1.615	.106
Confidence (z-scaled)	0.280 [–0.041, 0.600]	0.164	1.711	.087

Table 5: Logistic mixed-effects model predicting correctness by model and confidence for unary task.

	β [95% CI]	SE	z	p
Intercept (Model FishAudio)	–0.800 [–2.108, 0.508]	0.668	–1.199	.231
Model Hume	1.497 [–0.322, 3.316]	0.928	1.613	.107
Model MiniMax	0.129 [–1.688, 1.946]	0.927	0.139	.890
Model XTTS-v2	3.152 [1.240, 5.065]	0.976	3.230	**.001**
Confidence (z-scaled)	–0.485 [–0.778, –0.192]	0.150	–3.240	**.001**

The confidence-weighted unary AUROC was 0.65, indicating better-than-chance discrimination but weak separation of real and fake items (Figure 4). Participant-level analyses showed the same pattern (AUROC $M = 0.658$, 95% CI $[0.603, 0.713]$, $p < .001$; $d' = 0.692$, 95% CI $[0.492, 0.892]$, $p < .001$; and $c = 0.205$, 95% CI $[0.018, 0.393]$, $p = .032$). We interpret AUROC as supplementary to the accuracy and SDT results, as confidence level was measured on a 10-point Likert scale. The hit rate was 54.6% and the false-alarm rate was 25.4% (corresponding to the fake- and real-class rows of Figure 5).

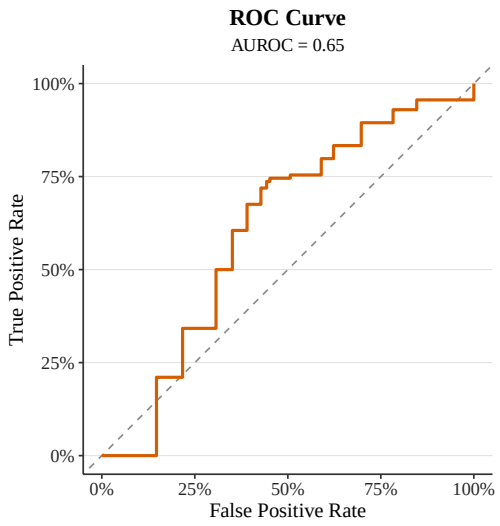


Figure 4: Unary-task detection analyses. ROC curve based on confidence-weighted real/fake responses.

4.5 Participants' Reasoning

The most frequently cited cues were global impressions of (un)naturalness (16.6%), phonetic detail (14.3%), prosody and delivery (11.8%), and speaker identity (10.2%); however, 30.4% of responses offered no stated cue and 18.6% appealed to subjective certainty, while recitation features were rare (0.8%). Because responses could receive multiple codes, percentages sum to more than 100%. In exploratory univariate mixed-effects models (Table 6), reasoning that appealed to subjective certainty was associated with lower accuracy (OR $= 0.53$, 95% CI $[0.33, 0.83]$, $p = .006$), although this effect narrowly misses significance after Holm correction across the nine cues. Concrete perceptual cues trended in the opposite direction (phonetic details: OR $= 1.69$, $p = .067$; naturalness: OR $= 1.64$, $p = .091$; prosody: OR $= 1.63$, $p = .11$), without reaching significance individually. Reasoning that referenced the speaker's identity was not associated with higher accuracy (OR $= 1.12$, $p = .65$). In a combined model adjusting for all cues, naturalness was a significant predictor (OR $= 1.92$, 95% CI $[1.07, 3.45]$, $p = .029$), phonetic and prosodic cues were marginal ($p = .054$, $p = .073$), and the subjective-certainty effect was attenuated ($p = .066$). No cue category significantly predicted confidence (all $p > .16$). We thus found no evidence that the type of reasoning listeners used related to how confident they felt, although the absence of significant effects does not establish that confidence and reasoning are independent. Given the number of tests, these patterns

should be regarded as hypotheses for confirmatory work.

5 Discussion

Three findings stand out in this study. Arabic listeners detect synthetic speech unreliably and their errors are systematically biased toward "real"; their confidence is poorly calibrated in the single-clip condition; and, in exploratory analysis, the kind of evidence they cite, rather than how certain they feel, appears to relate to accuracy.

Unreliable and systematically biased detection. Participants performed only moderately above chance, in line with a growing consensus that individuals cannot reliably detect speech deepfakes (Mai et al., 2023; Müller et al., 2022). Our contribution is to show that this conclusion extends to Arabic audio deepfakes, a setting absent from prior perception work, and to characterize *how* listeners err. Errors were not random: the signal-detection analysis revealed a bias toward judging audio as real (i.e., a tendency to judge audio as authentic; $c = 0.27$), and 45% of synthetic clips were accepted as real in the unary task. In the context of mis- and disinformation, this asymmetry is the more troubling direction of error, since it favors the uncritical acceptance of fabricated speech (Shoaib et al., 2023; Alanazi et al., 2025). That listeners defaulted to "real" is also consistent with evidence that contemporary voice clones are perceived as realistic, even if not yet hyperrealistic (Lavan et al., 2025): when synthesis is good enough, the absence of obvious artifacts is read as authenticity rather than as an open question.

Condition differences and the limits of metacognition. Once class imbalance is taken into account, accuracy was similar in the binary (62.9%) and unary (64.6% balanced) conditions, and the model-based difference was not significant. Absolute accuracy was lower than in Mai et al. (2023) in both formats, plausibly because our stimuli were generated largely with proprietary systems and presented in MSA, though stimulus set size and participant numbers also differ. The more notable contrast lies in metacognition. Confidence tracked accuracy in the binary condition but was inversely related to it in the unary condition, including on synthetic items alone. We found no evidence that confidence varied with the kind of reasoning a listener used. Because real-world exposure to suspect audio is almost always unary, this pattern is concerning, but our design does not allow us to attribute it to the absence of a reference clip.

What guides listeners' judgments. Analysis of participants' reasoning clarified the basis of their judgments. Reasoning that appealed to subjective certainty was associated with lower accuracy, whereas citing concrete perceptual cues (naturalness in particular, with phonetic and prosodic detail in the same direction) was associated with better performance, although these exploratory associations did not survive correction for multiple testing. This echoes Watson et al. (2021), who found that listeners' accuracy tracked surface features of the stimuli, and Groh et al. (2022), who showed that reliance on non-evidential cues can mislead human judgment. Model identity also mattered: in the unary task, clips from XTTS-v2 were detected more often than those from FishAudio and MiniMax, but the difference between proprietary and open-source systems was not significant. Detection difficulty therefore appears to be system-specific rather than a simple function of whether a model is proprietary or open-source, although with one speaker prompt per figure and three figures these model-level effects remain preliminary.

Interpreting the two conditions. The binary and unary conditions were designed to approximate two realistic scenarios rather than to isolate task format, and they differ along several dimensions at once: synthesis paradigm (anonymous-voice TTS vs. zero-shot cloning of recognizable figures), synthesis systems, recording sources (read-speech corpora vs. public addresses), clip duration (3–7 s vs. 7–10 s), speaker familiarity, and the presence of speaker images. Any of these could contribute to the observed differences in confidence and bias. For example, a recognizable voice and image may lend a clip credibility and raise confidence irrespective of its authenticity, which would produce both the "real" bias and the inverse confidence–accuracy relation independently of task format. Our results should therefore be read as descriptive of these two conditions. Isolating the effect of reference availability would require crossing task format with the same speakers, systems, and stimuli, and manipulating speaker images independently.

Directions for future work. Our results point to several directions rather than settled prescriptions. First, because we found no evidence that

Table 6: Univariate mixed-effects models of accuracy (logistic) and confidence (linear) for each reasoning cue, with random intercepts for participants and questions. ORs > 1 indicate higher accuracy when the cue is cited; β is in confidence points (0–10). n = responses citing the cue (multi-label, $N = 912$). p-values are uncorrected; no effect survives Holm correction across the nine cues.

		Accuracy (OR)		Confidence (β, pts)	
	n	OR [95% CI]	p	β [95% CI]	p
Audio quality	33	2.11 [0.66, 6.75]	.21	0.21 [-0.64, 1.06]	.63
Segmental and phonetic	130	1.69 [0.96, 2.95]	.067	0.04 [-0.46, 0.54]	.87
Prosodic and delivery	108	1.63 [0.90, 2.97]	.11	0.08 [-0.42, 0.58]	.75
Naturalness and humanness	151	1.64 [0.92, 2.90]	.091	-0.32 [-0.88, 0.24]	.27
Linguistic content	59	1.15 [0.61, 2.16]	.68	0.41 [-0.17, 1.00]	.17
Recitation features	7	0.81 [0.05, 13.28]	.88	0.11 [-2.18, 2.41]	.92
Speaker identity and familiarity	93	1.12 [0.68, 1.85]	.65	0.21 [-0.27, 0.69]	.39
Subjective certainty	170	0.53 [0.33, 0.83]	**.006**	-0.12 [-0.54, 0.29]	.56
Non-substantive	277	0.77 [0.54, 1.12]	.17	-0.21 [-0.59, 0.17]	.27

confidence tracked the cues listeners used and because it was inversely related to accuracy in the unary condition, interventions should not rely on listeners' self-assessed certainty. Prior work suggests that familiarizing listeners with deepfakes, whether as a concept or through examples, yields only modest gains (Mai et al., 2023; Watson et al., 2021), and whether cue-focused training helps Arabic listeners remains to be tested. Second, the association between citing naturalness, phonetic, and prosodic cues and higher accuracy suggests which features training materials could emphasize, a hypothesis that future intervention studies could evaluate directly. Third, aggregation of independent human judgments and human–machine collaboration have shown promise in other languages (Groh et al., 2022; Mai et al., 2023), with the caution that erroneous model outputs can themselves mislead a crowd (Groh et al., 2022). Evaluating automatic detectors on the same Arabic stimuli would clarify whether humans and machines err on the same items, and whether combining them improves detection.

Language-specific perception and the case for Arabic. Some listeners reasoned from distinctly Arabic phonological and register knowledge such as points of articulation (*makhārij*), diacritics (*tashkīl*). Although these cues were too infrequent to model statistically, their presence is theoretically important. They reflect language- and culture-specific competence that might not be represented in detectors trained on high-resource, monolingual benchmarks, and that speech synthesis systems are not optimized to reproduce. This supports the motivation for dedicated Arabic resources and detectors (Almutairi and Elgibreen, 2023; Awaad et al., 2024; Maged et al., 2025; Issa, 2026) and suggests that both human perception and machine detection in Arabic cannot be assumed to transfer from English. More broadly, it argues for treating perceptual studies as a complement to benchmark construction. Understanding which language-specific features human listeners exploit can inform the features that detectors, and public-awareness efforts, should prioritize for each linguistic community.

6 Conclusion

We presented a preliminary study of how Arabic listeners perceive audio deepfakes. Listeners performed moderately above chance in both conditions (binary: 62.9%; unary balanced accuracy: 64.6%). Errors were biased toward accepting synthetic speech as genuine, with 45% of cloned clips judged real, and confidence was inversely related to accuracy in the single-clip condition that most resembles real-world exposure, a pattern that our design cannot attribute to task format alone. Future work should compare human listeners with SOTA automatic detectors on the same stimuli, extend the study to dialectal Arabic, larger and more diverse listener populations, and larger sets of longer clips, and cross task format with matched stimuli.

Limitations

This study has some limitations. First, the sample is modest (38 listeners). Second, and more consequentially, the stimulus set is small: the TTS condition contains nine genuine–synthetic pairs and the cloning condition three speakers with twelve cloned variants. Although the mixed-effects models account for repeated measures, the small number of distinct items, speakers, and systems limits

generalization across speakers, linguistic content, and synthesis systems, and model-level effects in particular should be treated as preliminary. Third, the binary and unary conditions differ in more than task format, including the presence of speaker images in the unary condition, so differences between them cannot be attributed to format or reference availability. Fourth, the unary task is imbalanced (12 synthetic vs. 3 genuine clips), which constrains the precision of the false-alarm rate estimate. Fifth, stimuli and analysis were restricted to MSA; given the diglossic nature of Arabic (Ferguson, 1959), findings may not extend to the many spoken dialects, and we did not control for listeners' dialectal backgrounds or examine whether dialectal (mis)match between listener and stimulus affects detection. Finally, all clips were short (3–10 s), and longer stimuli may provide additional cues and yield different detection rates, particularly given conflicting prior findings on the effect of clip duration (Watson et al., 2021; Mai et al., 2023). We did not evaluate automatic detectors on the same stimuli, which would clarify whether human and machine errors coincide.

Ethics Statement

The study involved adult participants who took part voluntarily in an online listening experiment following approval from the authors' institutional Internal Review Board (IRB No.: STUDY-IRB-2025-1516). The voice-cloning stimuli were derived from publicly available YouTube recordings of well-known political figures, used solely to construct controlled experimental materials for studying human perception. The synthetic clips were created for research purposes only, were heard exclusively by consenting participants within the experiment, and were not published, distributed, or used to make any statement attributable to these individuals. We recognize that cloning the voices of real people is sensitive. Finally, ChatGPT was used to assist with English language editing. All such edits were reviewed and verified by the authors.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments, which helped improve this study; all remaining errors are our own. We also gratefully acknowledge the support provided by the Department of Linguistics and the School of Languages and Cultures at Purdue University.

References

Sami Alanazi, Seemal Asif, Antoinette Caird-daley, and Irene Moulitsas. 2025. Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses. *Human-Intelligent Systems Integration*, pages 1–23.

Zaynab M Almutairi and Hebah Elgibreen. 2023. Detecting fake audio of arabic speakers using self-supervised deep learning. *IEEE Access*, 11:72134–72147.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. *Advances in neural information processing systems*, 31.

Rewaa Awaad, Abdelrahman Hedya, Mariam Fadel, and Ahmed B Zaky. 2024. Real-time deepfake detection for ai-generated arabic speech. In *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 198–201. IEEE.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119.

Nawar Halabi and 1 others. 2016. Arabic speech corpus. *Oxford text archive core collection*.

Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. 2021. *Detection theory: A user's guide*. Routledge.

Elsayed Issa. 2026. AFAD-MSA: Dataset and models for Arabic fake audio detection. *Computation*, 14(1):20.

Zahra Khanjani, Gabrielle Watson, and Vandana P Janeja. 2023. Audio deepfakes: A survey. *Frontiers in Big Data*, 5:1001063.

Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. ISCA (the International Speech Communication Association).

Kartik Kuckreja, Parul Gupta, Injy Hamed, Thamar Solorio, Muhammad Haris Khan, and Abhinav Dhall. 2025. Tell me habibi, is it real or fake? *arXiv preprint arXiv:2505.22581*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Nadine Lavan, Mathilde Irvine, Victor Rosi, and Carolyn McGettigan. 2025. Voice clones sound realistic but not (yet) hyperrealistic. *PLOS ONE*, 20(9):e0332692.

Lantian Li, Yixiang Chen, Dong Wang, and Thomas Fang Zheng. 2017. A study on replay attack and anti-spoofing for automatic speaker verification. In *Proc. Interspeech*, pages 92–96.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2025. A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7):1–38.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and 1 others. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522.

Mohamed Maged, Alhassan Ehab, Ali Mekky, Besher Hassan, and Shady Shehata. 2025. Arfake: A multi-dialect benchmark and baselines for arabic spoof-speech detection. *Preprint*, arXiv:2509.22808.

Kimberly T Mai, Sergi Bray, Toby Davies, and Lewis D Griffin. 2023. Warning: Humans cannot reliably detect speech deepfakes. *Plos one*, 18(8):e0285333.

Cade Metz. 2019. Internet companies prepares to fight the 'deepfake' future. page B1. Accessed: 2025-09-01.

Nicolas M Müller, Karla Pizzi, and Jennifer Williams. 2022. Human perception of audio deepfakes. In *Proceedings of the 1st international workshop on deepfake detection for audio multimedia*, pages 85–91.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Mohamed R Shoaib, Zefan Wang, Milad Taleby Ahvanooey, and Jun Zhao. 2023. Deepfakes, misinformation, and disinformation in the era of frontier ai, generative ai, and large ai models. In *2023 international conference on computer and applications (ICCA)*, pages 1–7. IEEE.

John A Swets. 2014. *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. *arXiv preprint arXiv:2410.04527*.

N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. 2020. Introducing the VoicePrivacy Initiative. In *Proc. Interspeech 2020*, pages 1693–1697.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, and 1 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Gabrielle Watson, Zahra Khanjani, and Vandana P Janeja. 2021. Audio deepfake perceptions in college going populations. *arXiv preprint arXiv:2112.03351*.

Daniel Weiner and Nellie Lawrance. 2023. Regulating ai deepfakes and synthetic media in the political arena. New York Times, Accessed: 2025-09-01.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. 2015. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*, pages 2037–2041. International Speech Communication Association.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and 1 others. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, and 1 others. 2022. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, and 1 others. 2023. Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*.

A Audio Generation

Synthetic stimuli were produced under each model's default configuration. We made no changes to generation parameters such as temperature, stability, or similarity, and applied no post-hoc manipulation or editing to the resulting audio. This was a deliberate choice: relying on out-of-the-box settings approximates the conditions available to a typical user and ensures that detection difficulty reflects the systems' default output rather than tuning on our part.

For the binary (TTS) task, we used three proprietary models—ElevenLabs (https://elevenlabs.io/), MiniMax (https://www.minimax.io/), and ResembleAI (https://www.resemble.ai/)—accessed through their own web interfaces, without any custom code. For each genuine reference clip, we supplied the corresponding transcript to each system and generated a synthetic counterpart in the default voice/setting, producing the fake clips used in the pairwise discrimination trials.

For the unary (voice-cloning) task, we used four systems to clone each of the three reference speakers from a short sample of their authentic speech, again at default settings and with no manual editing of the output. Each system was prompted with the transcript of the corresponding genuine clip (Table 6), so all cloned variants reproduce the original wording of the speech. The two proprietary systems, Hume (https://www.hume.ai/) and MiniMax, were accessed through their web interfaces. The two open-source systems, FishAudio (https://huggingface.co/fishaudio/s1-mini) and XTTS-v2 (https://huggingface.co/coqui/XTTS-v2), were run from their publicly available HuggingFace model releases using a Python script, with default inference settings.

B Audio Stimuli Sources

Political Figure	Transcript	Source	Timestamp (dur.)
Abdel Fattah El-Sisi	شعب مصر العظيم السيدات والسادة الحضور الكريم. يظل المواطن المصري بطل هو روايتنا الوطنية.	YouTube[1]	0:00–0:10 (10s)
Mohammed bin Salman Al Saud	قامت هذه الدولة منذ ثلاثة قرون على مبادئ راسخة ترتكز على إعلاء الشريعة الإسلامية وإقامة العدل والشورى.	YouTube[2]	7:27–7:34 (7s)
Tamim bin Hamad Al Thani	ومازال أمامنا المزيد لنفعله لأن عزمنا على مواصلة تحقيق التقدم لبلدنا ولرفاهية شعبنا لا يقف عند حد.	YouTube[3]	1:35–1:43 (8s)

[1] https://www.youtube.com/watch?v=I2ZmBygvUoI
[2] https://www.youtube.com/watch?v=EPHziE2BZXY
[3] https://www.youtube.com/watch?v=ppHQJnHKuik

Table 7: Authentic source clips of political figures used for the voice-cloning experiments, showing the transcript, source URL, timestamp, and duration of each clip. The same transcripts were used to generate the cloned variants.

TTS	Audio ID	Transcript	Dur. (s)
ElevenLabs	common_voice_ar_24026124.mp3	وإن عليك لعنتي إلى يوم الدين	4
ElevenLabs	common_voice_ar_24034981.mp3	كان علىّ فعل هذا منذ أشهر خلت	4
ElevenLabs	common_voice_ar_24098642.mp3	متى سنقيم الحفلة	3
MiniMax	common_voice_ar_24033216.mp3	والوجه السابع	3
MiniMax	common_voice_ar_24024916.mp3	إننا نبحث عن منزل يضم مرآباً	4
MiniMax	common_voice_ar_24057081.mp3	والذين كفروا بآياتنا هم أصحاب المشئمة	4
ResembleAI	ARA_NORM__0077.wav	وبعثوا رسالة إلى الرئيس تتضمن مطالب لعودتهم	6
ResembleAI	ARA_NORM__0240.wav	مما يزيد الأمر تعقيداً	3
ResembleAI	ARA_NORM__0536.wav	دخلت الحضارة الإنسانية في طورٍ جديدٍ في تسعينيات القرن نفسه	7

Table 8: TTS stimuli used in the binary task, showing the synthesis model, audio ID, transcript, and duration of each clip.

C Reasoning Cue Coding Scheme

Table 9: Coding scheme for participants' free-text reasoning cues, with agreement per category between the two human coders.

Category	Definition	n	Agr. (%)	κ
Audio quality	Recording fidelity: noise, echo, clarity, quality.	33	95.4	0.52
Segmental and phonetic	Specific speech sounds: pronunciation, points of articulation, letters, diacritics.	130	96.2	0.86
Prosodic and delivery	Suprasegmentals/timing: intonation, stress, rate, pausing, elongation, rhythm, overlap, delivery.	108	97.1	0.87
Naturalness and humanness	Global human-vs-machine impression: natural/unnatural, robotic, "sounds AI", breathing, "too polished".	151	96.9	0.89
Linguistic content	Grammar, syntax, lexis, MSA correctness, fluency, coherence.	59	97.4	0.78
Recitation features	Recitation: verse pausing rules, stops, religious-register correctness.	7	99.7	0.73
Speaker identity and familiarity	Comparison to the known target speaker: "not his voice", recognizing the figure, accent/dialect of the person.	93	94.0	0.69
Subjective certainty	Subjective certainty or non-cue reasoning: confidence, "I think", "seems/feels real", guessing. May co-occur with other codes.	170	94.5	0.81
Non-substantive	Numeric-only, yes/no, or a reasoning with no stated cue.	277	87.1	0.67

D Full Tukey-adjusted Pairwise Contrasts Across Models

This appendix reports the complete set of Tukey-adjusted pairwise contrasts between speech synthesis models for both tasks, complementing the model-condition analyses in Section 4.3. Table 10 presents the contrasts among the binary (TTS) systems and Table 11 those among the unary (voice-cloning) systems. No binary contrast was significant. In the unary task, the reliable differences were that XTTS-v2 was detected significantly more often than FishAudio and MiniMax. In addition, Figure 5 shows the confusion matrices for binary and unary tasks.

Table 10: Tukey-adjusted pairwise contrasts among binary generated-item models.

Contrast	β [95% CI]	SE	z	p
ElevenLabs – MiniMax	0.055 [–1.973, 2.083]	0.865	0.063	.998
ElevenLabs – ResembleAI	1.390 [–0.627, 3.408]	0.861	3.885	.239
MiniMax – ResembleAI	1.335 [0.690, 3.361]	0.864	1.545	.270

Table 11: Tukey-adjusted pairwise contrasts among unary generated-item models.

Contrast	β [95% CI]	SE	z	p
FishAudio – Hume	–1.497 [–3.882, 0.887]	0.928	–1.613	.371
FishAudio – MiniMax	–0.129 [–2.510, 2.253]	0.927	–0.139	.999
FishAudio – XTTS-v2	–3.152 [–5.660, –0.645]	0.976	–3.230	.007
Hume – MiniMax	1.368 [–1.019, 3.756]	0.929	1.472	.454
Hume – XTTS-v2	–1.655 [–4.143, 0.833]	0.969	–1.709	.319
MiniMax – XTTS-v2	–3.023 [–5.537, –0.510]	0.978	–3.091	.011

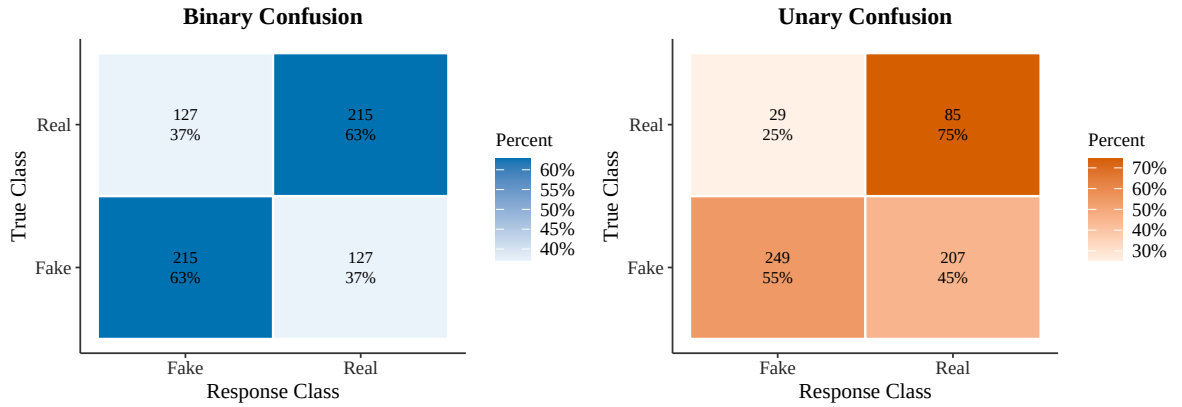
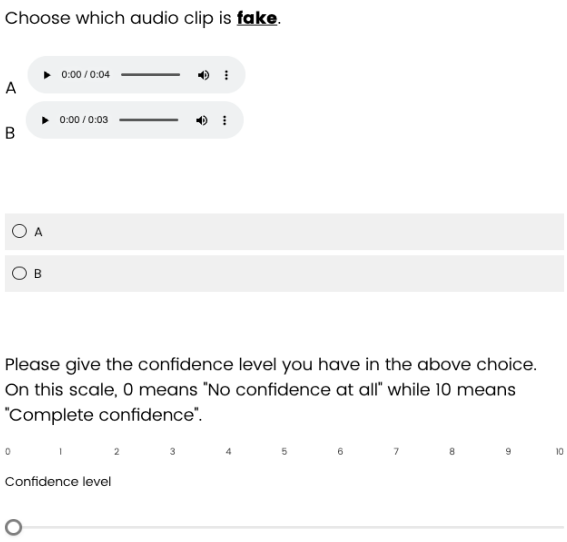
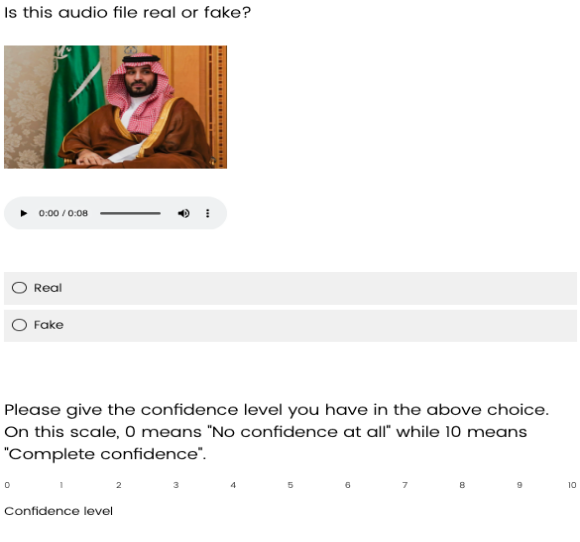


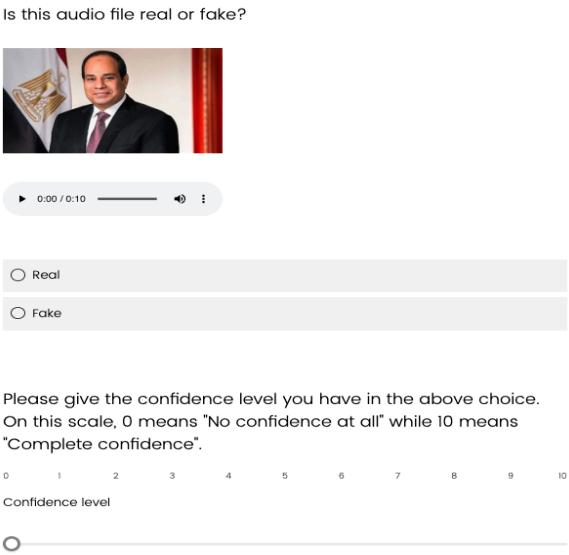
Figure 5: Confusion matrices for binary and unary tasks. Cells show counts and percentages. Note: For visualization only, each forced-choice binary response was expanded into two clip-level inferred classifications: the selected clip was coded as "fake" and the unselected clip as "real." Thus, the binary matrix contains 684 inferred classifications from 342 pairwise responses and should not be interpreted as independent clip-level judgments.



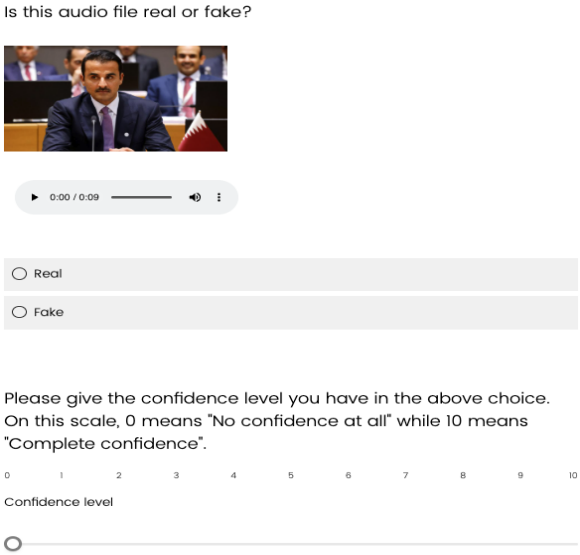
(a)



(b)



(c)



(d)

Figure 6: Example from the experiment. Panel (a) shows a binary text-to-speech task; panels (b)–(d) show unary voice-cloning for the three public figures whose voices were cloned. Each task presented the audio (and, for the voice-cloning trials, an image of the speaker), followed by the authenticity judgment, a 0–10 confidence rating, and a free-text prompt asking participants to explain their reasoning. The images of public figures are reproduced here solely to document the experimental interface for research purposes.