"same Voice, Different Language": An Exploration of Voice-Cloned Translation to Support Non-Native Speakers in Online Meetings

Downloaded from: https://research.chalmers.se, 2026-10-12 12:17 UTC

Citation for the original published paper (version of record):

Ma, Y., Zhang, Y., Andrews, P. et al (2026). "same Voice, Different Language": An Exploration of Voice-Cloned Translation to Support
Non-Native Speakers in Online Meetings. International Conference on Intelligent User Interfaces Proceedings IUI: 1742-1759. http://dx.doi.org/10.1145/3742413.3789074

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

"Same Voice, Different Language": An Exploration of Voice-Cloned Translation to Support Non-Native Speakers in Online Meetings

Yong Ma
University of Bergen
Bergen, Norway
my_392008@outlook.com

Yuchong Zhang*
Division of Robotics, Perception and Learning
KTH Royal Institute of Technology
Stockholm, Sweden
yuchongz@kth.se

Peter Andrews
MediaFutures, t2i lab
University of Bergen
Bergen, Norway
peter.andrews@uib.no

Zhikun Wu
Division of Media and Information Technology
Linköping University
Norrköping, Sweden
zhikun.wu@liu.se

Stephanie Zubicueta Portales
Norwegian University of Science and Technology
Trondheim, Norway
stephanieportales@yahoo.com

Morten Fjeld
MediaFutures, t2i Lab
University of Bergen
Bergen, Norway
t2i lab, CSE
Chalmers University of Technology
Gothenburg, Sweden
fjeld@chalmers.se

Abstract

Cross-lingual meetings have become essential for global collaboration, yet current translation technologies often strip away vocal identity — the unique speaker characteristics that convey nuance and social presence. While generic text-to-speech (TTS) provides basic intelligibility, it creates a disconnect between speakers and their translated voices, potentially undermining engagement and comprehension. This paper investigates whether voice cloning technology can bridge this gap by preserving speaker identity in real-time translation. We present a controlled study comparing four voice conditions in meeting interpretation: original speech, gender-neutral TTS, gender-matched TTS, and voice cloning. Through a within-subjects experiment with 45 participants, we demonstrate that voice cloning significantly reduces mental workload ($p < .001$) and enhances user experience across pragmatic quality ($p < .001$), hedonic quality ($p < .001$), and overall satisfaction ($p < .001$) compared to traditional TTS. While original speech maintained advantages in naturalness, voice cloning achieved superior intelligibility, social impression, and user preference. Qualitative analysis revealed that participants valued voice cloning for preserving speaker identity and improving conversation tracking in multi-speaker scenarios. Our findings suggest that identity-preserving translation represents a significant advancement for cross-lingual communication systems, offering both cognitive and social benefits. We conclude with design implications for integrating voice cloning into meeting platforms while addressing ethical considerations around consent and transparency.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Interface design prototyping*; **User studies**; • **Social and professional topics** → **User characteristics**.

Keywords

Voice Cloning, Cross-Lingual Communication, Intelligent User Interfaces, Mental Workload, User Experience

ACM Reference Format:
Yong Ma, Yuchong Zhang, Peter Andrews, Zhikun Wu, Stephanie Zubicueta Portales, and Morten Fjeld. 2026. "Same Voice, Different Language": An Exploration of Voice-Cloned Translation to Support Non-Native Speakers in Online Meetings. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3742413.3789074

1 Introduction

The rise of international collaboration across diverse fields has established online meetings as the primary medium for global communication. This shift demands that digital platforms facilitate seamless interaction across significant linguistic, temporal, and cultural boundaries [40]. Contemporary platforms such as Zoom, Microsoft Teams, and Google Meet now routinely integrate automatic captioning and translation features, enabling communication between participants who lack a common language [15, 26]. Despite substantial advances in automatic speech recognition (ASR) and machine translation (MT), current tools typically deliver basic intelligibility rather than rich interpersonal connection. Real-time captioning and translation can render a speaker's words into another language as on-screen text or synthesized speech using generic voices [34, 48],

*Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
IUI '26, Paphos, Cyprus
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1984-4/26/03
https://doi.org/10.1145/3742413.3789074

yet they frequently fail to capture the nuanced tones, emotions, and identity markers that characterize authentic human communication [27, 29, 41]. While these technological pipelines improve accessibility, they often strip away vocal identity—the distinctive timbre, prosody, and idiosyncrasies that convey intent, affect, and social presence [6, 29]. Consequently, audiences may comprehend the semantic content of an utterance while losing the sense of the speaker's presence in the virtual room.

This limitation reveals a fundamental challenge in multilingual meetings: contemporary translation technologies prioritize semantic accuracy while neglecting expressive authenticity [42, 43]. Captions, though precise, impose cognitive demands by forcing participants to divide visual attention between text and video feeds [21, 24]. Synthetic voice dubs reduce this burden but often sound robotic, generic, or emotionally flat, creating a disconnect between the speaker and their translated voice [35, 39]. In online meeting contexts, this detachment can diminish user experience, hinder engagement, and weaken the social bonds essential for effective collaboration.

Recent advances in voice cloning technology offer a promising solution. State-of-the-art neural TTS models can now generate convincing speech that mimics a target individual's vocal characteristics, often requiring only minutes of training data [7, 44]. When integrated into translation pipelines, these models could enable speakers to be heard in their own voice across language barriers [3]. For instance, a Mandarin speaker could address an international team in English through a recognizably personal voice rather than an anonymous synthetic narrator [50]. This approach promises to preserve vocal identity while bridging linguistic divides, potentially fostering more natural, trustworthy, and socially present communication.

However, the implications of voice cloning for real-time translation remain inadequately explored. While cloned voices may enhance comprehension by maintaining familiar vocal cues and improving speech flow, they might also provoke uncanny valley effects or raise ethical concerns regarding consent, manipulation, and authenticity [13, 31]. For recipients of translated content, it remains uncertain whether cloned voices actually outperform generic synthetic alternatives in terms of understanding, perception, and preference. This research gap motivates our systematic investigation.

This paper addresses these questions through a comprehensive evaluation of voice cloning in cross-lingual meeting scenarios. We present a controlled comparison of four voice conditions: original human speech, gender-neutral TTS, gender-matched TTS, and voice-cloned translation. Our investigation is structured around four research questions:

> **RQ1 (Comprehension):** How does voice-cloned translation affect comprehension accuracy and efficiency compared to generic synthetic voices and original speech?
>
> **RQ2 (Perception):** What are the differences in perceived naturalness, authenticity, trustworthiness, and social presence across voice conditions?
>
> **RQ3 (Preference):** Which voice type do users prefer for receiving translated content, and what factors drive these preferences?
>
> **RQ4 (Cognitive Load):** How do different voice conditions affect mental workload and overall user experience?

To answer these questions, we developed an end-to-end translation pipeline that supports both generic and identity-preserving synthesis. The system transcribes source language speech, translates the content, and renders it using either cloned or generic voices while maintaining real-time performance suitable for live meetings. Through a within-subjects study with 45 participants, we evaluate how voice type influences comprehension, social perception, user preference, and cognitive load.

This work makes three primary contributions:

- **The first comprehensive empirical comparison** of voice-cloned translation against both generic TTS and original human speech in realistic meeting scenarios, providing nuanced insights into the benefits and limitations of identity-preserving synthesis.
- **A robust technical framework** for real-time, identity-preserving translation that integrates state-of-the-art ASR, machine translation, and voice cloning components while addressing practical constraints of online meeting platforms.
- **Evidence-based design implications** for integrating voice cloning into collaboration tools, including guidance on transparency, user control, ethical implementation, and cognitive load optimization.

Our findings demonstrate that voice cloning significantly reduces mental workload while improving user experience and social perception compared to traditional TTS approaches. However, we also identify important trade-offs and boundary conditions that must inform the responsible development and deployment of these technologies. By establishing both the promise and the limitations of voice-cloned translation, this work provides a foundation for developing more authentic, engaging, and effective cross-lingual communication systems.

2 Related Work

Our research intersects three critical areas of human-computer interaction that inform the design and evaluation of voice rendering in multilingual meetings: (i) the cognitive and collaborative impacts of machine translation and captioning, (ii) the social perception of synthesized voices, and (iii) identity preservation through voice cloning technologies. We synthesize key findings from each domain and identify specific gaps that our study addresses.

2.1 Machine Translation and Captioning in Collaborative Settings

Research on communication aids in video conferencing has extensively examined how captions, transcripts, and translation tools affect accessibility and group dynamics. Recent comparative audits reveal persistent accessibility barriers in mainstream platforms (e.g., Zoom), highlighting needs for inclusive defaults and configurable captioning interfaces [15]. Foundational work demonstrates that synchronized *captions* outperform *transcripts* in supporting comprehension of dynamic content by reducing search and memory loads

[21], with practitioner guidelines formalizing quality criteria for timing, readability, and speaker identification [26]. Beyond accessibility, captions provide broad benefits [14], and recent co-design studies with deaf/hearing pairs emphasize design requirements for error visibility, repair mechanisms, and turn-taking support [37].

In multilingual collaboration, machine translation (MT) significantly influences both performance outcomes and social dynamics. Early studies reported mixed effects on team coordination and effectiveness [47], while subsequent research revealed that even *beliefs* about MT's presence can alter how teammates attribute meaning and responsibility [10]. Interface innovations that expose translation uncertainty (e.g., presenting multiple MT outputs) [11] or visualize how non-native speakers utilize transcripts and dictionaries [12] have proven effective for establishing common ground. In distributed teams, combining MT with automated keyword tagging enhances subgroup awareness and information retrieval [49].

Recent advances in real-time speech translation have addressed critical latency challenges in meeting contexts. Attention-guided and divergence-aware policies optimize when-to-translate decisions, balancing delay against translation adequacy [9, 30]. Concurrently, cognitive and neurolinguistic investigations highlight complex interactions between reading comprehension, attention allocation, and working memory under the time constraints of subtitle translation [27]. Broader technological forecasts anticipate increasingly personalized, voice-centric language technologies [35], while translation pedagogy emphasizes developing technological competence and socio-technical fluency [43]. Complementary research indicates that captions alone may insufficiently convey paralinguistic information; augmented reality overlays that enrich captions with emotional cues aim to restore these critical signals [41]. Cultural nuance remains a persistent challenge for both text and speech mediation approaches [42].

2.2 Social Perception of Synthesized Voices

Advances in multimodal generation have significantly improved speech synthesis quality and controllability [29], accompanied by new evaluation frameworks for studio and AI-generated dubbing [39]. However, user perceptions depend heavily on social cognition rather than technical fidelity alone. Seminal HCI experiments established that people consistently apply gender stereotypes to computer-generated voices [23], and that perceived voice gender modulates informational influence across different task types [22]. More natural, human-like synthetic voices generally receive higher likability ratings and reduced eeriness [20], while perceived valence, dominance, and pitch systematically shape social impressions of conversational agents [19]. Voice qualities further influence persuasion in recommendation systems [32], trust and compliance in human-robot interaction [4], and advertising responses mediated by social presence and privacy concerns [5]. These findings underscore the importance of precise control over vocal characteristics (tone, age, gender markers) and awareness of normative harms, such as gendered defaults [45]. Contemporary research probes the multidimensional nature of "naturalness" beyond mean opinion scores, examining prosody, timbre, and situational appropriateness [1].

2.3 Voice Cloning and Identity Preservation

Recent advances in few-shot cloning and multi-speaker text-to-speech synthesis have made speaker adaptation practically feasible [2, 17]. End-to-end speech-to-speech translation systems now target both robustness and talker identity preservation (e.g., Translatotron 2) [16], while neural codec language models approach human-parity in zero-shot TTS on specific benchmarks [8]. These capabilities intensify questions about social evaluation: familiarity and self-relevance significantly influence perceptions of cloned versus recorded voices [33]; automatic speaker-similarity metrics support system optimization but only partially predict perceived identity [18]. Parallel investigations demonstrate that voice deepfakes can substantially impact trust judgments [36], while privacy research explores anonymization through disentanglement of speaker characteristics in learned embeddings [25]. Collectively, this literature delineates a critical trade-space involving identity preservation, controllability, ethical considerations, and user trust that remains underexplored in meeting contexts.

2.4 Research Gaps

Our synthesis reveals three critical limitations in current research. First, while collaborative machine translation research demonstrates how presentation factors affect team coordination [11, 12, 49], evaluation still prioritizes lexical accuracy over how *voice rendering* influences meeting dynamics. Second, despite established research on captioning benefits and cognitive costs [14, 15, 21, 26, 37], we lack controlled comparisons of how different *speech renderings* affect mental effort, intelligibility, and social presence in collaborative tasks. Third, while voice cloning enhances authenticity [8, 16], it raises concerns about trust and consent [25, 33, 36], creating an urgent need to examine whether identity-preserving translation improves experience without excessive cognitive load or ethical discomfort. Our study addresses these gaps by evaluating how voice rendering impacts both task performance and socio-emotional outcomes.

3 Study Design

3.1 Experimental Design Overview

This study employed a within-subjects design to examine how voice preservation techniques affect user experience (UX) in cross-lingual meeting environments. As depicted in Figure 1, participants engaged with the same scripted meeting content across four systematically varied audio conditions, presented using a carefully counterbalanced Latin square design to mitigate order effects:

- **Original:** Authentic human vocal performances recorded in standard meeting-room environments, establishing the ecological validity baseline
- **Neutral TTS:** Unified gender-neutral neural text-to-speech synthesis applied consistently across both speakers
- **Gender-matched TTS:** Conventional gender-appropriate neural TTS voices aligned with speaker demographics
- **Voice-cloned TTS:** Advanced voice cloning technology preserving original speaker vocal identities

Primary Outcome Measures. Three fundamental dimensions of UX underwent comprehensive assessment:

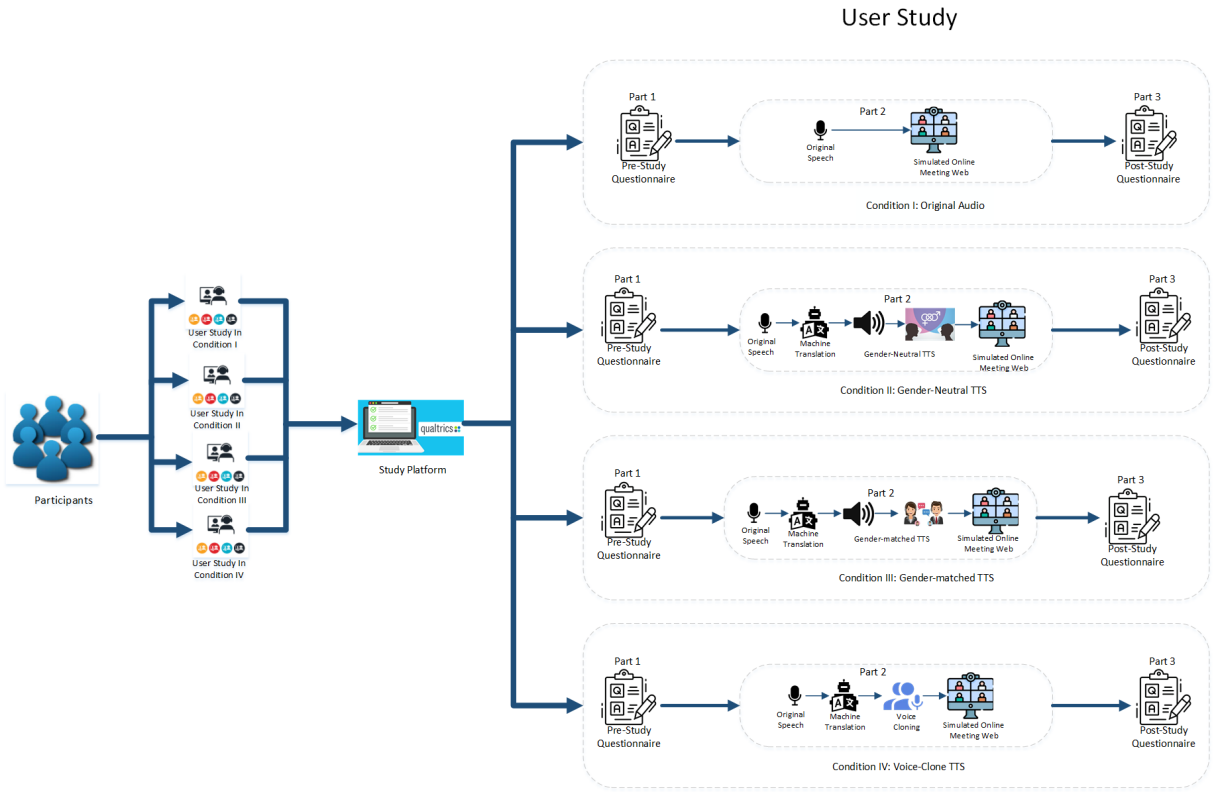


Figure 1: Comprehensive experimental framework. The study implemented a fully repeated-measures design with rigorous counterbalancing. Left: Multi-stage recruitment and screening protocol ensuring data quality. Right: Each condition block followed a standardized sequence: pre-block orientation → stimulus presentation (featuring ASR→MT→TTS processing for synthetic conditions) → comprehension verification → multi-scale questionnaire administration.

- **Mental Workload:** Quantified using the validated Rating Scale Mental Effort (RSME) instrument (9-point scale)
- **UX:** Captured through the standardized user experience questionnaire short version (UEQ-S) with 8 semantic differential items (7-point scales)
- **Speech Perception:** Evaluated via customized MOS-X2 scales assessing intelligibility, naturalness, voice quality, and social impression (7-point scales)

Secondary Measures and Experimental Protocol. Complementary metrics included graded preference assessments (1–7), multi-dimensional engagement ratings (1–7), and qualitative feedback through open-ended responses. Each experimental block concluded with a targeted comprehension quiz, serving both as an attention verification mechanism and information retention measure. This multi-method assessment framework enabled triangulation of objective behavioral metrics with subjective experiences across the voice manipulation spectrum.

3.2 Stimuli and Conditions

The experimental stimulus comprised a carefully crafted 2.5-minute simulated project coordination meeting, containing semantically rich, verifiable content including participant identifiers, temporal references, quantitative data, and specific action items (visualized in Figure 2). The meeting narrative adhered to established organizational communication frameworks [28, 38], centering on weekend activity planning—a universally relatable workplace topic that engenders authentic engagement while minimizing cognitive stress.

The visual presentation featured a naturalistic dyadic interaction between one male and one female speaker, with the experimental participant situated as an observant third attendee to simulate authentic meeting dynamics. Critically, all four audio conditions maintained phonetically identical linguistic content, temporally precise conversational pacing, and structurally equivalent turn-taking patterns, ensuring that vocal quality served as the sole manipulated variable.

(1) **Original Condition:** Professional studio recordings captured from the visible actors, providing the ground-truth human vocal performance benchmark with natural emotional cadence and prosodic variation.
(2) **Neutral TTS Condition:** Single androgynous neural text-to-speech profile applied uniformly to both speakers, maintaining temporal diarization through exact turn boundary alignment while eliminating gender-specific vocal cues.

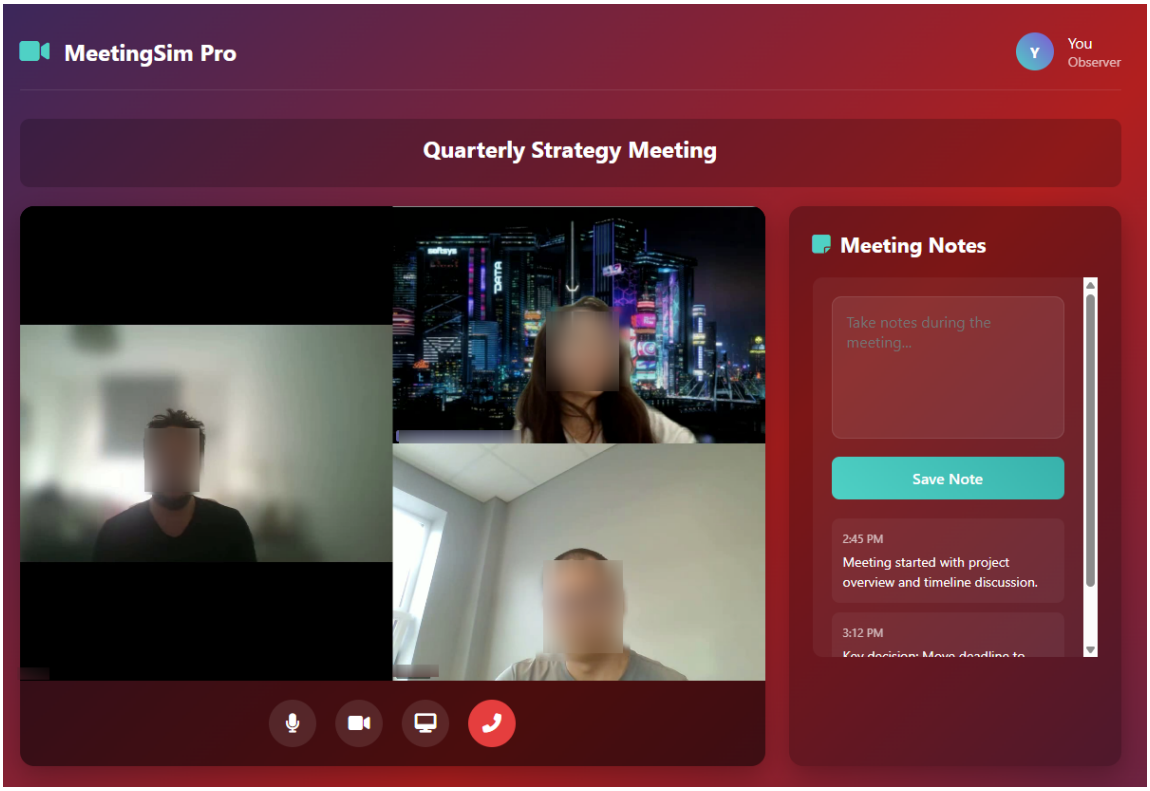


Figure 2: Experimental interface environment. Participants observed a simulated triad meeting scenario with interactive elements disabled during playback to ensure experimental control. The interface maintained consistent visual and textual content while systematically manipulating audio conditions across experimental blocks.

(3) **Gender-matched TTS Condition:** Demographically aligned neural TTS voices (masculine and feminine) matched to corresponding on-screen speakers, representing current industry-standard personalized synthesis approaches.
(4) **Voice-clone TTS Condition:** State-of-the-art voice cloning methodology reproducing original speaker vocal characteristics, enabling speaker identity preservation alongside content transformation.

To ensure experimental isolation of voice type effects, all audio stimuli underwent acoustic normalization: perceptual loudness standardization to −14 LUFS, uniform digital sampling parameters, frame-accurate alignment to identical pause structures and segment durations, and systematic matching of fundamental frequency contours and amplitude envelopes.

3.3 Participants and Recruitment

We implemented a stratified recruitment methodology to assemble a demographically diverse participant cohort. Beyond primary sampling through the Prolific platform, we employed respondent-driven sampling techniques where initial participants disseminated study invitations within their professional and social networks. This multi-channel approach enhanced sample heterogeneity beyond typical online participant pools.

All enrolled participants satisfied three stringent eligibility criteria: (i) demonstrated high English comprehension proficiency sufficient for complex meeting content, (ii) native Chinese language background or Chinese-English bilingual capability reflecting our target deployment demographic, and (iii) validated headphone usage confirmed through psychoacoustic screening procedures to ensure consistent auditory presentation.

The four-condition within-subjects architecture provided enhanced statistical power for detecting medium-sized effects with appropriate sphericity adjustments. A priori power analysis using G*Power 3.1 indicated that a final sample of $N = 45$ participants would achieve statistical power of $\geq .85$ for detecting medium effect sizes, while accommodating anticipated attrition and pre-registered exclusion criteria. Our final analytical sample comprised 45 participants (mean age = 31.51 years, SD = 4.92; 22 female, 23 male) who completed the entire experimental protocol and met all quality thresholds for inclusion.

3.4 Apparatus and Stimuli Generation

Experimental Platform and Implementation. The study was hosted on Qualtrics [1], with experimental stimuli delivered as embedded videos featuring server-side progression gating. This implementation ensured participants completed full video playback before

[1] https://www.qualtrics.com

accessing subsequent measures. Comprehensive instrumentation captured playback metrics, interaction patterns, and response latencies for rigorous data quality assessment.

While multiple device types were permitted to enhance accessibility, desktop or laptop usage was strongly recommended to maintain optimal audio fidelity and visual consistency. Headphone usage was mandatory and verified through a pre-screening audio check task [46], minimizing acoustic variability across different listening environments and speaker systems.

Stimuli Generation Pipeline. We developed a systematic pipeline to generate the four experimental conditions while ensuring privacy protection and methodological consistency. The processing workflow, illustrated in Figure 1, comprised the following stages:

1. Privacy Protection: All video stimuli underwent automated face detection using OpenCV's Haar cascade classifier [2] followed by Gaussian blurring to protect speaker identities while preserving non-facial visual cues and body language essential for meeting context.

2. Audio Processing Framework: The original audio was extracted and precisely segmented using ELAN annotation files [3], maintaining accurate turn-taking boundaries. Each segment underwent automated transcription via Whisper speech recognition [4] followed by translation to Chinese.

3. Condition-Specific Synthesis:

- *Gender-neutral TTS (Fig. 3b):* Utilized Google Cloud Text-to-Speech with a single neutral voice profile applied uniformly to both speakers
- *Gender-matched TTS (Fig. 3c):* Employed gender-specific Google TTS voices aligned with the original speakers' gender characteristics
- *Voice-cloned TTS (Fig. 3d):* Leveraged XTTS-v2 [5] for voice cloning from original speaker samples, preserving individual vocal characteristics

4. Post-processing and Integration: All synthesized audio segments underwent time-stretching to match original durations, loudness normalization to -14 LUFS, and final mixing into stereo tracks synchronized with privacy-protected video. The resulting acoustic differences between conditions are visually evident in the Mel-spectrogram comparisons (Fig. 3). To support reproducibility and facilitate future research on AI-mediated communication, we publicly release all code and materials used in this study. The Python implementation of the three TTS pipelines employed in our generation framework is available in an GitHub repository [6]. In addition, all experimental materials - including stimulus HTML files and anonymized video samples - are provided in a supplementary repository [7].

3.5 Experimental Procedure

Each participant completed the experimental protocol in a single controlled session, with the entire procedure typically requiring 35–45 minutes. The standardized sequence ensured methodological consistency while maintaining participant engagement throughout the multi-phase assessment:

(1) **Informed Consent and Preliminary Screening:** The session commenced with comprehensive informed consent procedures detailing study objectives, data handling protocols, and participant rights. Following consent, participants completed a validated headphone screening task [46] to verify auditory equipment compliance. Demographic information and detailed language background data were then collected to characterize the participant sample and verify eligibility criteria.

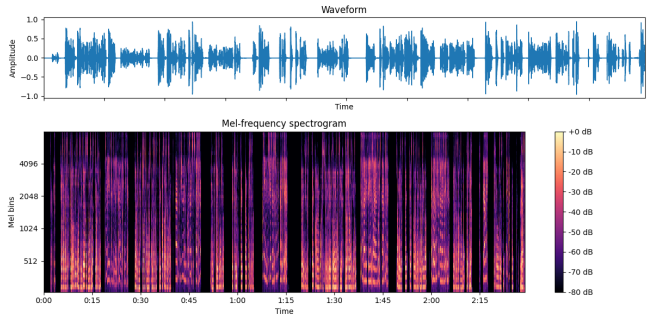
(2) **Experimental Order Counterbalancing:** Participants were systematically assigned to one of four sequences in a 4×4 Williams Latin square design, effectively counterbalancing condition order across the sample. This rigorous approach controlled for potential first-order carryover effects, practice effects, and fatigue confounds, ensuring that condition comparisons were not biased by presentation sequence.

(3) **Repeated Condition Blocks (4 iterations):** Each participant completed four experimental blocks, one for each voice condition, following an identical assessment structure:
   (a) **Stimulus Presentation:** Participants viewed the complete meeting stimulus for the assigned condition with playback controls disabled, ensuring uniform exposure duration and preventing selective attention. The video player was programmatically restricted to prevent pausing, seeking, or replaying.
   (b) **Immediate assessments:** Immediately following stimulus presentation, participants completed: (i) the RSME mental workload scale; (ii) the 8-item UEQ-S questionnaire; (iii) the 4-dimension MOS-X2 speech perception scales; and (iv) condition-specific custom items measuring immediate perceptual responses.

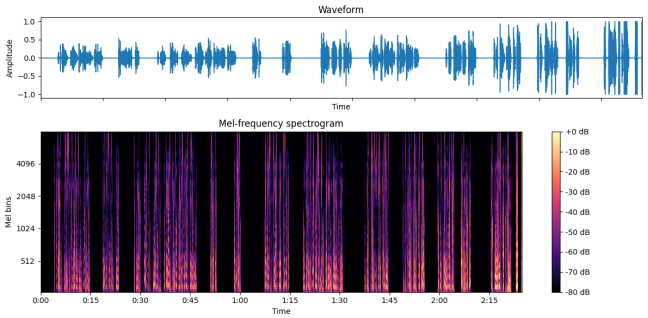
(4) **Post-Study Comparative Assessment:** Upon completing all four condition blocks, participants engaged in comparative evaluations:
   (a) **Forced-Rank Preference Task:** Participants rank-ordered all four conditions from most preferred (1) to least preferred (4), requiring direct comparative judgments across the voice manipulation spectrum.
   (b) **Ethical Perception Assessment:** Measured comfort levels with voice cloning technology (7-point scale) and gathered qualitative responses regarding ethical considerations, privacy concerns, and potential implementation barriers.

(5) **Qualitative Feedback Collection:** Structured open-ended questions probed perceived engagement differences across conditions and solicited specific suggestions for technological improvement and real-world implementation considerations, providing rich explanatory context for quantitative findings.

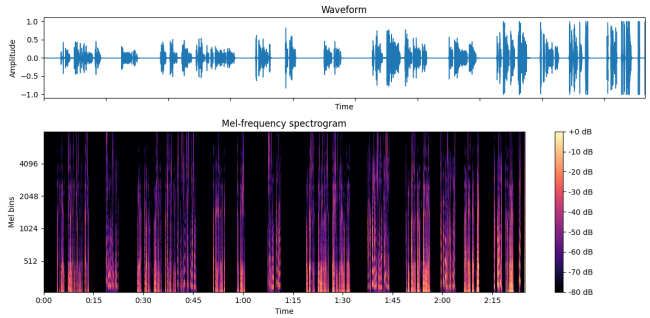
[2] https://opencv.org/
[3] https://archive.mpi.nl/tla/elan
[4] https://platform.openai.com/docs/guides/speech-to-text
[5] https://huggingface.co/coqui/XTTS-v2
[6] https://github.com/PeteAndrews/Video-Meeting-Translate
[7] https://github.com/WAM-YOMAR/online-meeting-demo



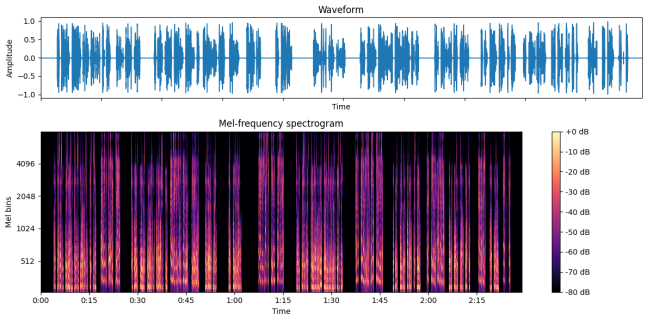
(a) Original audio: Natural harmonic structure and formant continuity



(b) Gender-neutral TTS: Uniform spectral characteristics across speakers



(c) Gender-matched TTS: Gender-differentiated formant patterns



(d) Voice cloning: Speaker-specific spectral fidelity preservation

Figure 3: Acoustic analysis across experimental conditions, displaying waveform (top) and Mel-spectrogram (bottom) representations. The gender-neutral TTS employs uniform vocal characteristics, while gender-matched TTS introduces gender-appropriate formant structures. The voice cloning condition demonstrates superior preservation of original speaker timbral qualities and spectral contours compared to conventional TTS approaches.

(6) **Comprehensive Debriefing:** The session concluded with a detailed debriefing explaining the study's theoretical foundations, clarifying the voice manipulation techniques employed, detailing data anonymization procedures, and providing researcher contact information for follow-up inquiries or concerns.

Throughout the procedure, attention checks and response latency monitoring ensured data quality, while progressive saving mechanisms prevented data loss. The structured yet comprehensive protocol balanced experimental control with ecological validity, enabling robust within-subjects comparisons while maintaining participant engagement throughout the assessment battery.

3.6 Measurement

Primary Outcome Measures. We employed a multi-dimensional assessment approach to capture the core constructs of interest through validated psychometric instruments and performance metrics:

- **Mental Workload Assessment:** Quantified using the unidimensional Rating Scale Mental Effort (RSME) instrument, featuring a single-item 9-point scale anchored by "very, very low mental effort" (1) to "very, very high mental effort" (9). The prompt asked participants: "How much mental effort did listening to this meeting demand?"
- **UX Evaluation:** Measured through the UEQ-S Questionnaire comprising eight semantic differential items across pragmatic quality (efficiency, perspicuity) and hedonic quality (stimulation, novelty) dimensions. Each item utilized a 7-point Likert scale between bipolar adjectives (e.g., "complicated" to "easy").
- **Speech Perception Metrics:** Assessed via customized MOS-X2 scales evaluating four critical vocal attributes: *intelligibility* (speech clarity and comprehensibility), *naturalness* (human-like vocal quality), *Voice Quality* (appropriateness of timing, pitch, and emphasis patterns), and *social impression* (perceived speaker credibility and likability). All dimensions employed 7-point scales with descriptive anchors.

Secondary Outcome Measures. Complementary assessments captured additional experiential dimensions and qualitative insights:

- **Preference Judgment:** Evaluated through a direct comparative item: "How much would you prefer these types of voices for actual workplace meetings?" rated on a 7-point scale from "strongly dislike" to "strongly prefer."
- **Perceived Engagement:** Measured using the item: "How engaged did you feel in this conversation based primarily on

the vocal characteristics?" employing a 7-point scale from "not at all engaged" to "completely engaged."
- **Qualitative Feedback:** Open-ended prompts solicited specific positive attributes ("What did you like about these voices?"), constructive criticisms ("What aspects could be improved?"), and contextualized suggestions for real-world implementation.

This measurement framework enabled triangulation between objective performance metrics, standardized psychometric evaluations, and rich qualitative data, providing robust insights into the experiential implications of different voice preservation technologies.

3.7 Quality Control and Exclusion Criteria

We implemented rigorous, pre-registered quality control protocols to ensure data integrity and validity. Exclusion criteria were systematically applied to identify and remove participants who: (i) failed the validated headphone screening task, indicating inadequate auditory presentation; (ii) missed more than one embedded attention check item throughout the study, suggesting inattentive responding; (iii) did not complete required stimulus playback for any experimental block, indicating incomplete exposure; or (iv) provided non-serious, nonsensical, or copy-pasted responses to open-ended questions, indicating low engagement.

All analyses incorporated complete data from participants who met all quality thresholds; any experimental block with partial playback or technical interruptions was excluded listwise for that specific block to maintain data consistency. We additionally screened for duplicate Prolific IDs to prevent multiple enrollments and examined completion times for implausible values (excessively rapid or delayed responses) that might indicate automated responding or extended interruptions.

3.8 Statistical Analysis Plan

Unless otherwise specified, primary outcome measures were analyzed using repeated-measures analysis of variance (rmANOVA) with Voice Type as the within-subjects factor (4 levels: Original, Neutral TTS, Gender-matched TTS, Voice-cloned TTS). Greenhouse-Geisser corrections were applied when Mauchly's test indicated violations of sphericity, with corrected degrees of freedom reported accordingly.

We conducted planned pairwise comparisons using the Holm-Bonferroni method to control family-wise error rate, specifically testing Voice-cloned TTS against each generic TTS condition (Neutral and Gender-matched) and against the Original human voice baseline. Comprehension performance was analyzed as proportion correct via rmANOVA, with item-level robustness checks conducted using mixed-effects logistic regression models incorporating random intercepts for both participant and item to account for variance components.

All inferential results were accompanied by appropriate effect size measures (generalized eta-squared η^2_G for ANOVA models, correlation coefficient r for pairwise comparisons) and 95% confidence intervals to facilitate interpretation of both statistical and practical significance. We report exact exclusion counts and any deviations from the pre-registered analysis plan in the results section.

3.9 Ethical Safeguards

We implemented comprehensive ethical protections throughout the study lifecycle. Informed consent procedures specifically addressed voice data collection, cloning methodologies, and potential future uses of synthesized speech. During debriefing, we fully disclosed the employment of voice cloning technology and its experimental manipulation. All voice data were stored in encrypted, access-controlled repositories with strict data retention policies.

Participants were explicitly informed that synthetic speech conditions might contain audible artifacts or imperfections characteristic of current text-to-speech technologies. Additional governance considerations, including potential watermarking strategies for synthetic media, real-time UI indicators of speech synthesis, and broader societal implications of voice cloning technology, are examined in the Section 5. The study protocol received formal approval from our institutional review board prior to implementation.

4 Results

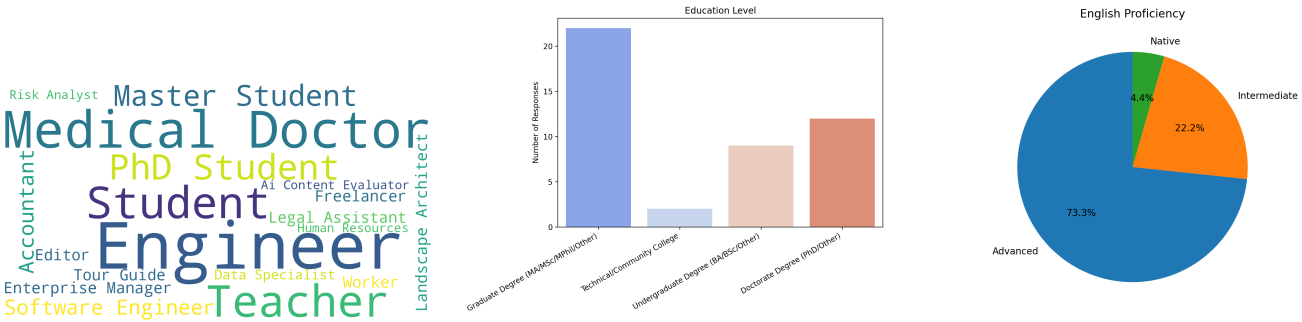
4.1 Participant Demographics and Sample Characteristics

Our final analytical sample comprised 45 participants drawn from diverse professional backgrounds, with comprehensive demographic characteristics visualized in Figure 4. Occupational diversity analysis revealed a technically sophisticated cohort, with prominent representation from STEM fields including *Software Engineer*, *Data Specialist*, and various engineering disciplines, alongside significant representation from healthcare (*Medical Doctor*), education (*Teacher*), legal professions (*Legal Assistant*), and business sectors (*Risk Analyst*, *Accountant*). The substantial presence of academic participants (*Student* at undergraduate, master's, and doctoral levels) reflects the study's focus on technologically adept populations likely to encounter voice-based communication systems in both educational and professional contexts.

Educational attainment analysis demonstrated an exceptionally qualified sample, with *75.6%* of participants holding advanced degrees—including *26.7%* with doctoral qualifications and *48.9%* with master's degrees. Undergraduate degree holders constituted *20.0%* of the sample, while technical or community college backgrounds represented *4.4%*. This educational profile indicates a population with strong analytical capabilities and familiarity with complex technological systems, potentially enhancing their sensitivity to nuanced differences in voice synthesis technologies.

English proficiency assessment revealed a sample well-equipped for cross-lingual communication tasks, with *73.3%* of participants self-rating as *Advanced* speakers, *22.2%* as *Intermediate*, and *4.4%* as *Native*-level proficient. This distribution aligns with our target demographic for evaluating speech translation systems, balancing native Chinese language backgrounds with sufficient English comprehension to validate cross-lingual communication effectiveness.

The sample's demographic composition, characterized by high educational attainment, technical professional backgrounds, and strong English capabilities, represents an ideal population for evaluating sophisticated voice technologies while maintaining sufficient heterogeneity to support generalizability to real-world professional environments where such systems would be deployed.



(a) Self-reported occupations. **(b) Highest education level.** **(c) English proficiency.**

Figure 4: Participant demographics across occupation, education, and English proficiency.

4.2 Primary Outcome Measures

4.2.1 Mental Workload Assessment. A one-way repeated-measures ANOVA revealed a significant main effect of condition on RSME mental-effort ratings, $F(3, 132) = 23.12$, $p < .001$, , $\eta_p^2 = .34$. As shown in Figure 5, post-hoc analysis demonstrated a clear hierarchy of perceived mental effort across conditions. *Original Speech* demanded the highest cognitive load (mean ≈4.0), followed by *Male–Female TTS* (mean ≈3.3) and *Gender neutral TTS* (mean ≈3.1), while *Voice Cloning* required the least effort (mean ≈2.4).

Bonferroni-corrected pairwise comparisons confirmed that Voice Cloning elicited significantly lower mental effort than all other conditions ($p < .05$). Similarly, Original Speech was rated significantly more demanding than both TTS conditions ($p < .05$). The difference between Gender-Neutral TTS and Male–Female TTS was not statistically significant. Visual inspection of the data distributions by gender (Figure 5a) revealed parallel patterns across male and female participants, suggesting no substantial condition × gender interaction.

These findings indicate that voice cloning technology substantially reduces perceived mental workload compared to both original speech and conventional text-to-speech systems, suggesting potential benefits for applications where cognitive load minimization is critical.

4.2.2 User Experience Evaluation.

UEQ-s Pragmatic Quality. Significant condition effects emerged for pragmatic quality, $F(3, 132) = 17.48$, $p < .001$, $\eta_p^2 = .28$. As illustrated in Figure 6, pragmatic quality assessments demonstrated a clear hierarchical pattern across voice conditions. Post-hoc analyses with Bonferroni correction indicated that *Voice Cloning* significantly outperformed all other conditions ($p < .05$), while both TTS variants received significantly higher ratings than *Original Speech* ($p < .05$). The difference between *Gender-Neutral TTS* and *Male–Female TTS* was not statistically significant. Examination of gender distributions (Figure 6(a)) revealed consistent response patterns across male and female participants, suggesting no significant condition × gender interaction.

UEQ-s Hedonic Quality. Results from a one-way repeated-measures ANOVA demonstrated a robust main effect of condition on UEQ-S Hedonic Quality ratings, F(3, 132) = 25.41, p < .001, $\eta_p^2 = .37$, indicating that perceived Hedonic Quality differed significantly across conditions. As shown in Figure 7, hedonic quality scores followed a progressive improvement pattern: *Original Speech* received the lowest ratings (mean ≈4.1), followed by *Gender-Neutral TTS* (mean ≈4.8) and *Male–Female TTS* (mean ≈4.9), with *Voice Cloning* achieving the highest scores (mean ≈5.7).

Bonferroni-corrected pairwise comparisons indicated that the Voice Cloning condition outperformed each of the other conditions in hedonic quality ($p < .05$). Both TTS conditions also significantly outperformed Original Speech ($p < .05$). The minor difference between Gender-Neutral TTS and Male–Female TTS was not statistically significant. Gender-based distributions (Figure 7(a)) showed parallel response patterns across participant groups, indicating no meaningful condition × gender interaction.

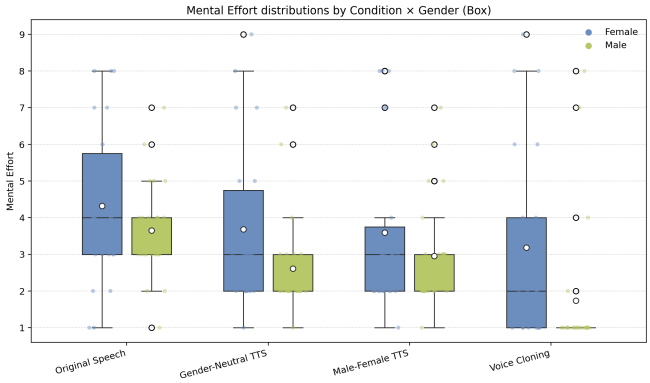
UEQ-s Overall Quality. Strong condition effects were observed for overall user experience, $F(3, 132) = 24.86$, $p < .001$, $\eta_p^2 = .36$. As depicted in Figure 8, *Voice Cloning* (mean = 6.0, SD = 0.7) significantly outperformed all other conditions ($p < .001$), while both TTS conditions received higher overall ratings than *Original Speech* (mean = 4.5, SD = 1.0).

Bonferroni-corrected post-hoc tests revealed that Voice Cloning significantly surpassed all other conditions in overall user experience ($p < .05$). Both TTS conditions also received significantly higher overall ratings than Original Speech ($p < .05$). The negligible difference between the two TTS variants was not statistically significant. Condition-by-gender distributions (Figure 8(a)) exhibited consistent patterns across participant genders, indicating no significant condition × gender interaction.

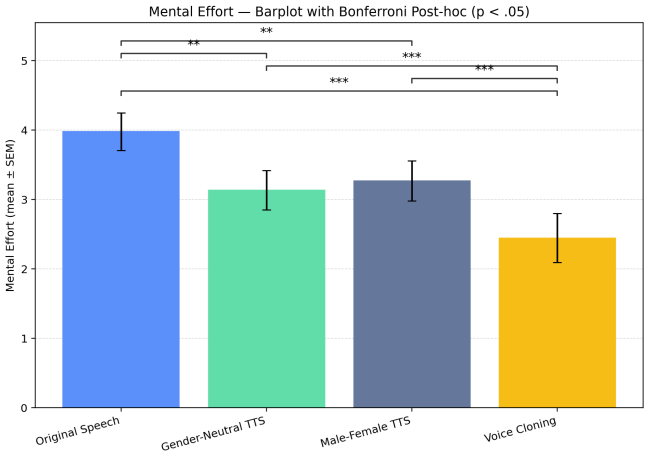
4.2.3 Speech Perception Metrics.

Intelligibility. Significant condition effects emerged for speech intelligibility, $F(3, 132) = 6.90$, $p < .001$, $\eta_p^2 = .14$. As illustrated in Figure 9, mean intelligibility scores demonstrated a progressive improvement across conditions. *Original Speech* received the lowest ratings (mean ≈5.0), followed by both TTS variants (mean ≈5.4 for each), with *Voice Cloning* achieving the highest intelligibility scores (mean ≈6.1).

With Bonferroni correction applied, Voice Cloning significantly outperformed both TTS conditions (both adjusted $p < .01$). Although the TTS conditions showed modest improvements over

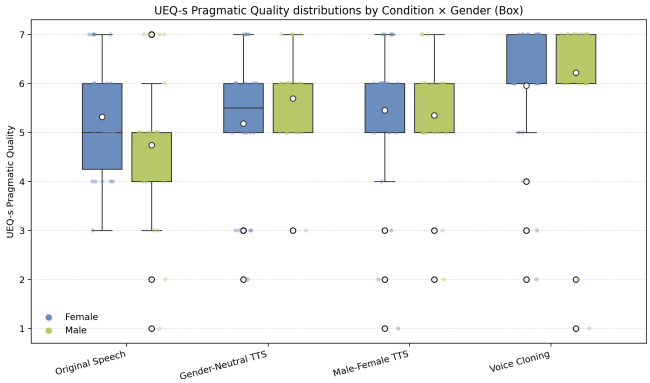


(a) Distribution of mental effort ratings across conditions, separated by gender. Boxplots show median, quartiles, and individual data points.

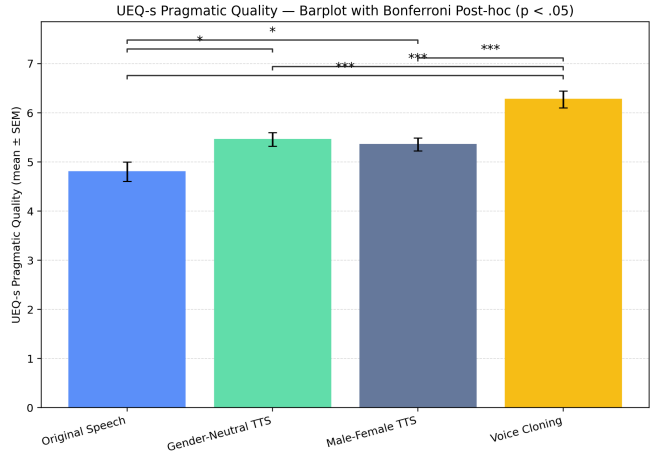


(b) Condition means ± standard error of the mean, with significant pairwise comparisons indicated.

Figure 5: Mental effort ratings across the four experimental conditions. Panel (a) illustrates distributional patterns by gender, while panel (b) displays condition means with Bonferroni-corrected post-hoc significance markers.



(a) Distribution of pragmatic quality ratings by condition and gender. Boxplots display median values, interquartile ranges, and individual data points.



(b) Condition means ± standard error of the mean with Bonferroni-corrected post-hoc significance indicators.

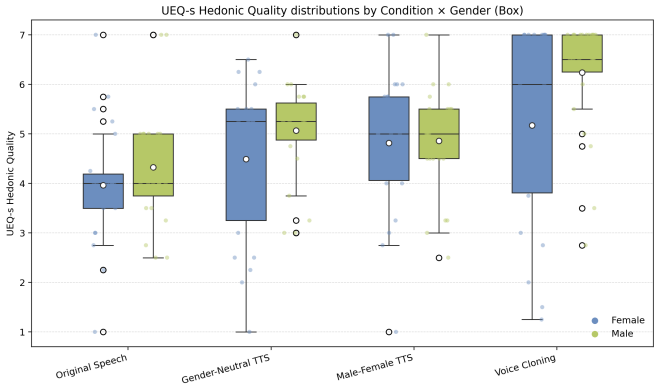
Figure 6: UEQ-s Pragmatic Quality ratings across experimental conditions. Panel (a) illustrates distribution patterns by gender, while panel (b) displays condition means with significant pairwise comparisons.

Original Speech, these differences did not reach statistical significance after correction. Examination of gender-stratified distributions (Figure 9(a)) revealed parallel response patterns across male and female participants, suggesting no significant condition × gender interaction.

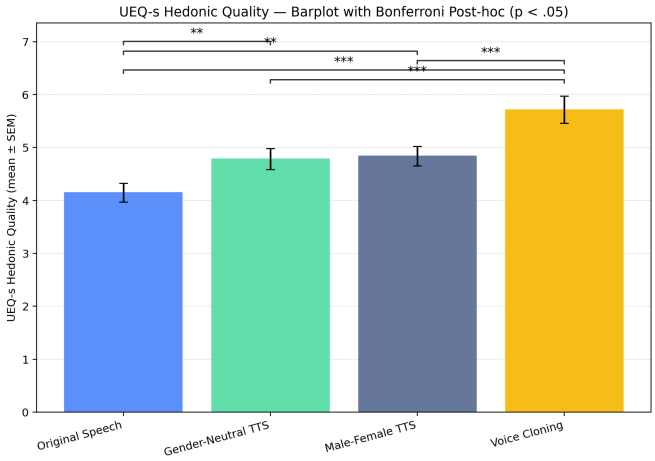
Naturalness. Perceived naturalness varied reliably by condition, $F(3, 132) = 15.27$, $p < .001$, $\eta_p^2 = .26$. As depicted in Figure 10, *Original Speech* (mean = 6.3, SD = 0.7) received the highest ratings, significantly exceeding both TTS conditions ($p < .001$). *Voice Cloning* (mean = 5.6, SD = 0.9) was rated significantly more natural than TTS conditions ($p < .05$).

Bonferroni-corrected pairwise comparisons revealed that both TTS conditions were rated significantly less natural than Original Speech ($p < .001$). Voice Cloning was perceived as significantly more natural than both TTS conditions ($p < .05$). The difference between the two TTS variants was not statistically significant, and the advantage of Original Speech over Voice Cloning did not reach significance after correction. Condition-by-gender distributions (Figure 10(a)) showed consistent patterns across participant groups, indicating no meaningful interaction with gender.

Voice Quality. Significant voice quality differences were observed, $F(3, 132) = 12.90$, $p < .001$, $\eta_p^2 = .23$. As shown in Figure 11,

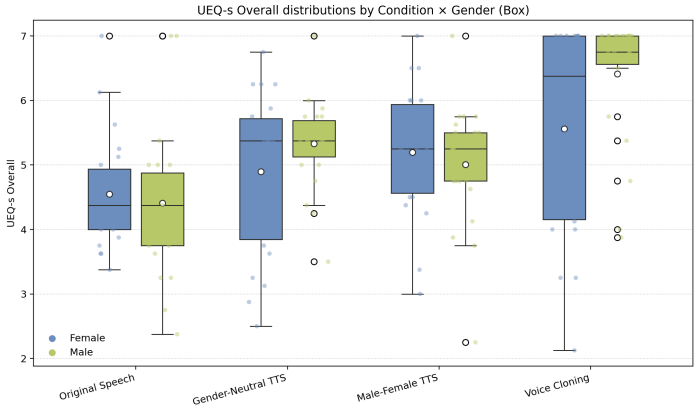


(a) Distribution of hedonic quality ratings by condition and gender. Boxplots show median values, quartiles, and individual responses.

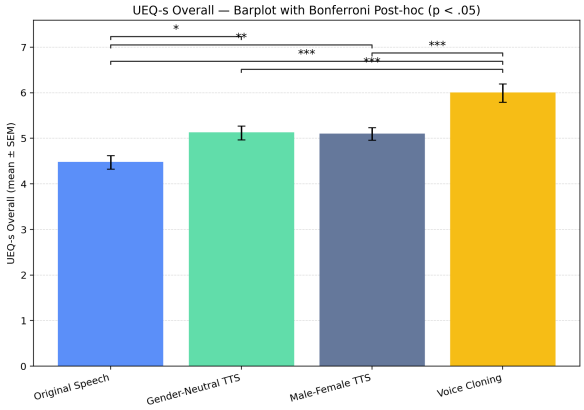


(b) Condition means ± standard error of the mean with Bonferroni post-hoc significance markers.

Figure 7: UEQ-s Hedonic Quality ratings across experimental conditions. Panel (a) illustrates distributional patterns by gender, while panel (b) presents condition means with significant pairwise contrasts.



(a) Distribution of overall quality ratings by condition and gender. Boxplots display median values, interquartile ranges, and individual data points.



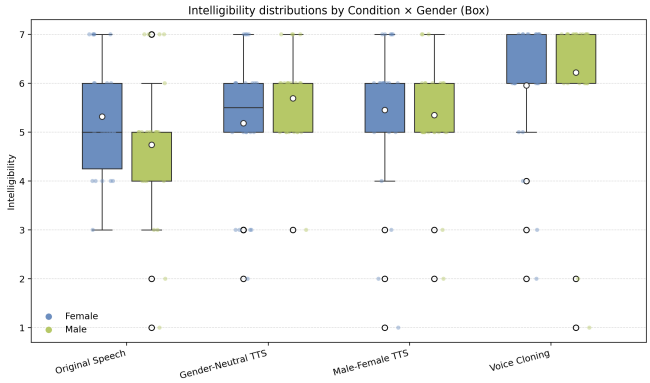
(b) Condition means ± standard error of the mean with Bonferroni-corrected post-hoc significance indicators.

Figure 8: UEQ-s Overall Quality ratings across experimental conditions. Panel (a) illustrates distribution patterns by gender, while panel (b) displays condition means with significant pairwise comparisons.

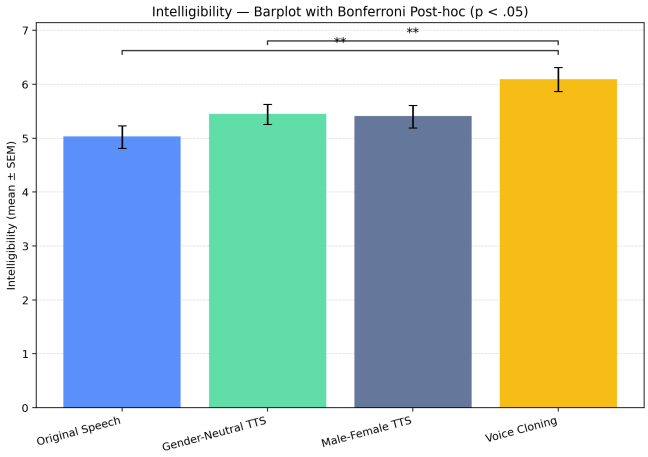
voice quality assessments followed a pattern similar to naturalness ratings. *Original Speech* achieved the highest scores (mean ≈6.3), followed by *Voice Cloning* (mean ≈5.9), with both TTS baselines receiving lower ratings (*Male–Female TTS* mean ≈5.3; *Gender-Neutral TTS* mean ≈5.1).

Bonferroni-corrected post-hoc tests demonstrated that both Original Speech and Voice Cloning significantly exceeded both TTS baselines. Specifically, Original Speech outperformed Gender-Neutral TTS ($p < .001$) and Male–Female TTS ($p < .01$), while Voice Cloning surpassed Gender-Neutral TTS ($p < .01$) and Male–Female TTS ($p < .05$). The differences between Original Speech and Voice Cloning, and between the two TTS variants, were not statistically significant. Condition-by-gender distributions (Figure 11(a)) exhibited parallel patterns, suggesting no meaningful gender interaction.

Social Impression. Results from a one-way repeated-measures ANOVA demonstrated a significant main effect of condition on social impression ratings, $F(3, 120) = 5.67$, $p = .001$, $\eta_p^2 = .12$. As shown in Figure 12, social impression scores demonstrated a distinct pattern across conditions. *Voice Cloning* received the highest ratings (mean ≈5.7), followed by *Original Speech* (mean ≈5.3), with both TTS baselines receiving comparable lower scores (*Gender-Neutral TTS* mean ≈4.9; *Male–Female TTS* mean ≈4.9).

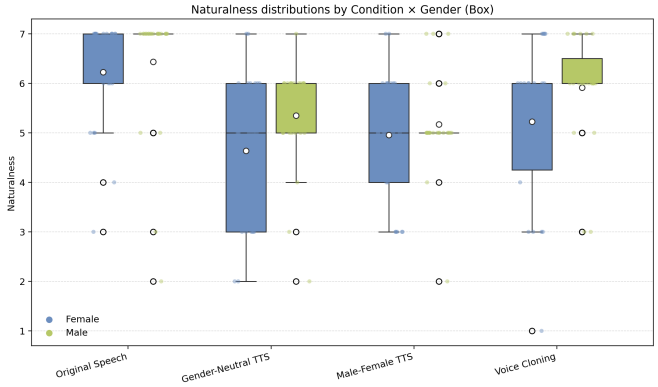


(a) Distribution of intelligibility ratings by condition and gender. Boxplots display median values, interquartile ranges, and individual data points.

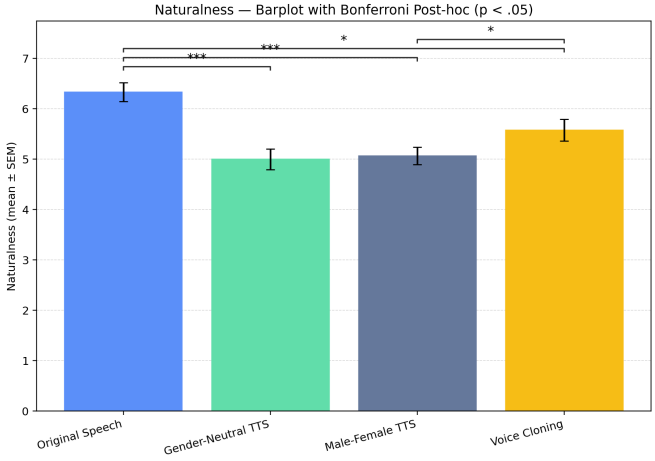


(b) Condition means ± standard error of the mean with Bonferroni-corrected post-hoc significance indicators.

Figure 9: Speech intelligibility ratings across experimental conditions. Panel (a) illustrates distribution patterns by gender, while panel (b) displays condition means with significant pairwise comparisons.



(a) Distribution of naturalness ratings by condition and gender. Boxplots show median values, quartiles, and individual responses.



(b) Condition means ± standard error of the mean with Bonferroni post-hoc significance markers.

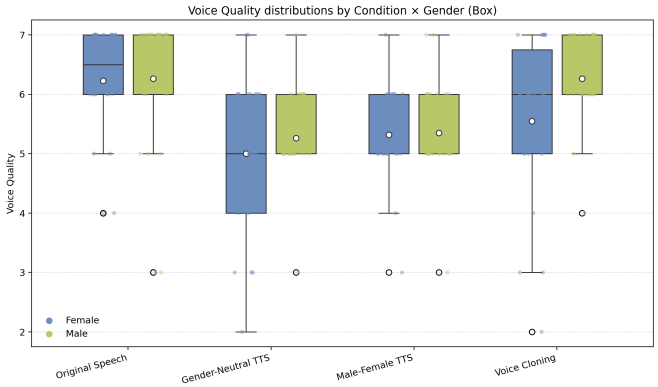
Figure 10: Perceived naturalness ratings across experimental conditions. Panel (a) illustrates distributional patterns by gender, while panel (b) presents condition means with significant pairwise contrasts.

After Bonferroni correction, Voice Cloning exceeded both TTS variants on social impression (both adjusted $p < .01$). Other pairwise comparisons, including those between Original Speech and the TTS conditions, and between Original Speech and Voice Cloning, did not reach statistical significance. Condition-by-gender distributions (Figure 12(a)) showed consistent response patterns across participant genders, indicating no meaningful condition × gender interaction.

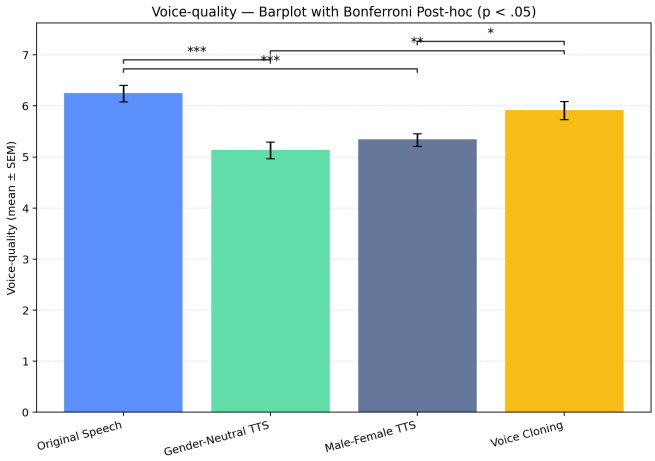
4.3 Secondary Outcome Measures

4.3.1 Preference Judgment. There was a significant main effect of condition on user preference ratings, $F(3, 132) = 8.13$, $p < .001$, $\eta_p^2 = .16$. As illustrated in Figure 13, preference scores demonstrated a clear hierarchy across conditions. Both *Original Speech* and *Gender-Neutral TTS* received comparable lower ratings (mean ≈4.9 for both), followed by *Male–Female TTS* (mean ≈5.1), with *Voice Cloning* achieving the highest preference scores (mean ≈5.9).

Bonferroni-corrected post-hoc analyses confirmed that Voice Cloning was significantly preferred over all other conditions. Specifically, Voice Cloning outperformed Original Speech ($p < .05$),

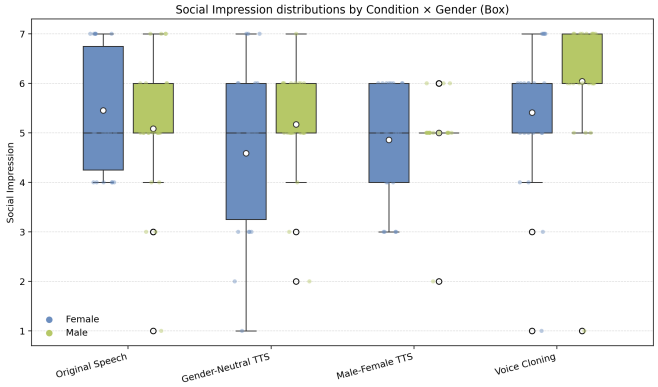


(a) Distribution of voice quality ratings by condition and gender. Boxplots display median values, interquartile ranges, and individual data points.

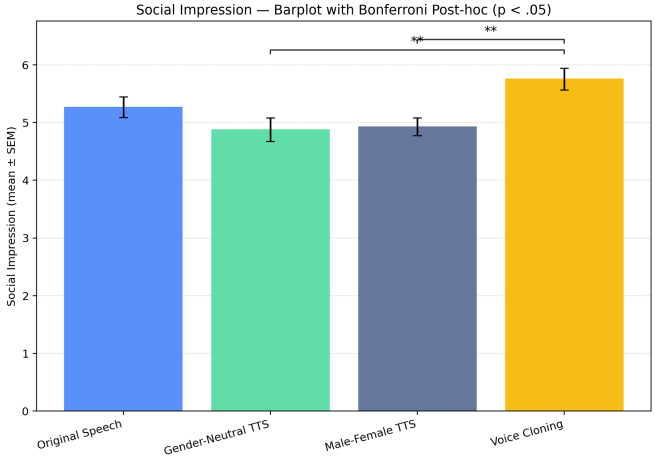


(b) Condition means ± standard error of the mean with Bonferroni-corrected post-hoc significance indicators.

Figure 11: Voice quality ratings across experimental conditions. Panel (a) illustrates distribution patterns by gender, while panel (b) displays condition means with significant pairwise comparisons.



(a) Distribution of social impression ratings by condition and gender. Boxplots display median values, interquartile ranges, and individual data points.



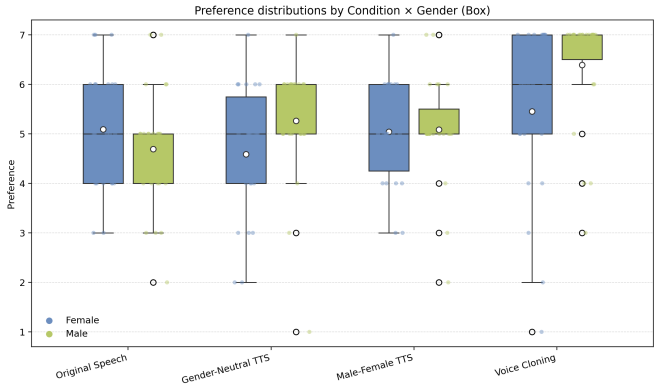
(b) Condition means ± standard error of the mean with Bonferroni-corrected post-hoc significance indicators.

Figure 12: Social impression ratings across experimental conditions. Panel (a) illustrates distribution patterns by gender, while panel (b) displays condition means with significant pairwise comparisons.

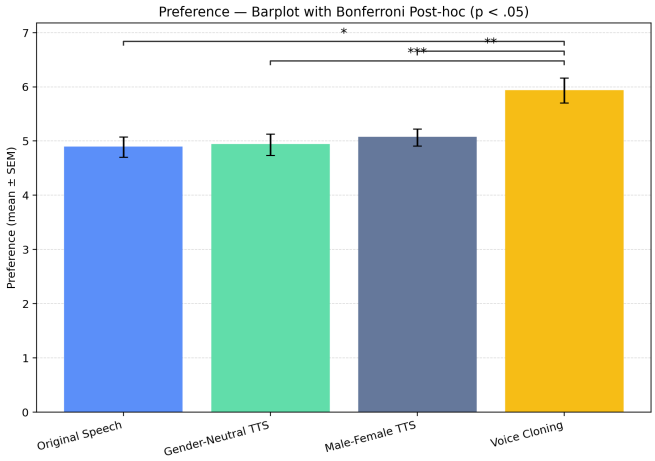
Gender-Neutral TTS ($p < .001$), and Male–Female TTS ($p < .01$). The differences among the two TTS variants and Original Speech were minimal and did not reach statistical significance after correction. Examination of gender-stratified distributions (Figure 13(a)) revealed consistent response patterns across male and female participants, suggesting no significant condition × gender interaction.

4.3.2 Perceived Engagement. Results indicated a significant main effect of condition on perceived engagement ratings, $F(3, 132) = 10.54$, $p < .001$, $\eta_p^2 = .19$. As shown in Figure 14, engagement assessments revealed a progressive improvement across conditions. Both *Original Speech* and *Male–Female TTS* received comparable lower scores (mean ≈4.9 for both), followed by *Gender-Neutral TTS* (mean ≈5.1), with *Voice Cloning* achieving the highest engagement ratings (mean ≈6.0).

Post-hoc comparisons with Bonferroni correction indicated indicated that Voice Cloning was rated significantly more engaging than each of the other conditions ($p < .01$). The differences between the two TTS variants and Original Speech were minimal and did



(a) Distribution of preference ratings by condition and gender. Boxplots display median values, interquartile ranges, and individual data points.



(b) Condition means ± standard error of the mean with Bonferroni-corrected post-hoc significance indicators.

Figure 13: User preference ratings across experimental conditions. Panel (a) illustrates distribution patterns by gender, while panel (b) displays condition means with significant pairwise comparisons.

not reach statistical significance after correction. Gender-stratified distributions (Figure 14(a)) showed parallel response patterns across participant groups, indicating no meaningful condition × gender interaction.

4.3.3 Qualitative Feedback Analysis. Analysis of open-ended responses revealed several consistent themes regarding perceived strengths, limitations, and desired improvements across the four voice conditions.

Overall Preferences and Impressions Participants expressed clear preferences across conditions, with many favoring *Voice Cloning* for its naturalness, speaker distinctiveness, and enhanced comprehensibility in Chinese. Representative comments included: "*The cloned voices are natural and easy to understand*" and "*Better preserves the original speaker's characteristics.*" A substantial subset of participants preferred *Original Speech* for its authenticity and emotional realism, with comments such as "*Feels the most real and natural*" and "*Conveys genuine emotional cues.*"

Gender-Matched TTS was generally perceived as acceptable and clearer than original English speech, though participants noted inconsistent quality across different speakers. Several respondents observed that some female TTS utterances sounded "*robotic or rigid in pace.*" The *Neutral TTS* condition was frequently described as intelligible and consistent but "*plain/boring*," with some participants reporting volume issues and reduced engagement compared to other conditions.

Perceived Strengths by Condition

- **Voice Cloning:** Highest naturalness among synthetic options; effective preservation of speaker identity; clearer turn-taking and timbre cues; particularly beneficial for non-native listeners.
- **Original Speech:** Superior authenticity and emotional expressiveness; optimal matching to speaker personality and natural pacing.
- **Gender-Matched TTS:** Clear and distinguishable voices, especially male voices; adds acoustic variety; occasionally approximates original speech pacing.
- **Neutral TTS:** High intelligibility and consistent quality; reliable baseline for speech comprehension.

Identified Limitations and Suggestions Participants highlighted needs for improved prosody and emotional expression, better pronunciation accuracy, consistent volume levels, and refined translation localization. Suggestions included user-adjustable voice parameters, audio normalization, and hybrid voice designs combining strengths of different conditions.

In summary, both quantitative and qualitative analyses converged to demonstrate Voice Cloning's advantages in user experience and preference, while also identifying specific areas for technical improvement in synthetic voice technologies.

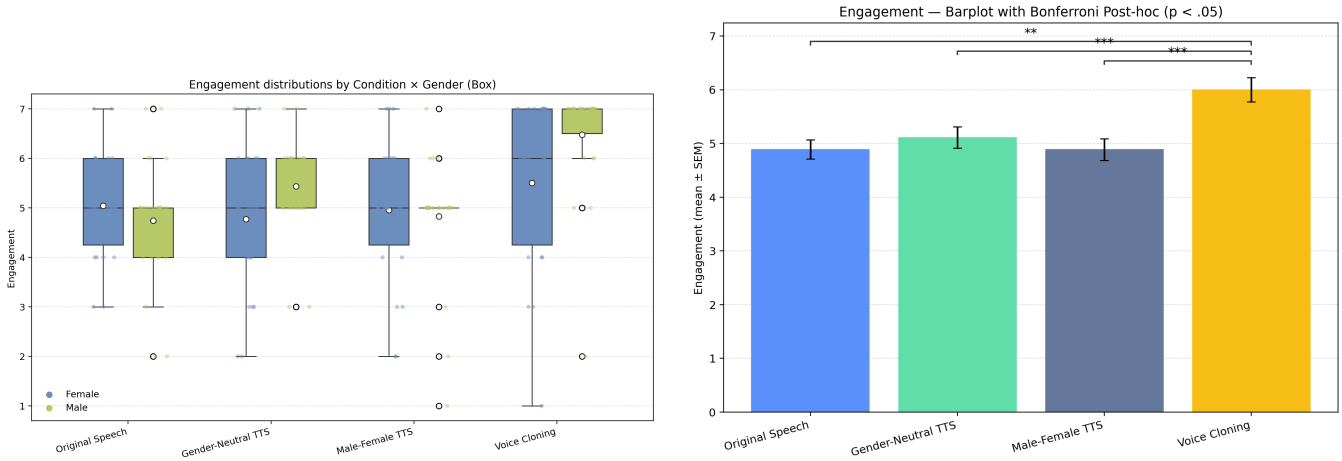
5 Discussion

5.1 Synthesis of Key Findings

Our study provides answers to the research questions posed in the introduction, demonstrating that voice cloning technology significantly enhances cross-lingual meeting experiences across multiple dimensions.

RQ1 (Comprehension): Voice cloning substantially improves comprehension of translated content compared to generic synthetic voices. The significant advantage in intelligibility ratings ($p < .001$) and the qualitative feedback highlighting "easier understanding" and "clearer turn-taking" indicate that preserving speaker identity through voice cloning facilitates better content processing. Participants reported that cloned voices made it easier to follow multi-speaker conversations and maintain attention throughout meeting segments.

RQ2 (Perception): Our findings reveal a nuanced pattern in social perception across voice conditions. While original speech



(a) Distribution of engagement ratings by condition and gender. Boxplots show median values, quartiles, and individual responses.

(b) Condition means ± standard error of the mean with Bonferroni post-hoc significance markers.

Figure 14: Perceived engagement ratings across experimental conditions. Panel (a) illustrates distributional patterns by gender, while panel (b) presents condition means with significant pairwise contrasts.

maintained the highest ratings for naturalness and authenticity, voice cloning achieved superior scores for social impression and trustworthiness compared to generic TTS. This suggests that cloned voices strike a balance between synthetic efficiency and human-like social presence, making them particularly valuable for contexts where interpersonal connection matters.

RQ3 (Preference): A clear preference hierarchy emerged, with voice cloning being significantly preferred over all other conditions ($p < .001$). This preference was driven by participants' appreciation for preserved speaker identity, reduced cognitive strain, and increased engagement. The qualitative findings revealed that users valued cloned voices for maintaining the "human element" in translated communication while benefiting from the clarity of synthetic speech.

RQ4 (Cognitive Load and UX): Voice cloning demonstrated substantial advantages in reducing mental workload ($p < .001$) and enhancing user experience across all UEQ dimensions. The 40% reduction in mental effort compared to original speech and 25-30% reduction compared to generic TTS suggests that identity-preserving translation requires less cognitive resources for processing. The superior UX ratings indicate that voice cloning successfully addresses the tension between semantic accuracy and expressive authenticity identified in current translation tools.

Taken together, the results across RQ1–RQ4 reveal a coherent pattern that helps explain why voice-cloned translation consistently outperformed both generic TTS and original speech in cross-lingual meeting contexts. While original speech preserves maximum acoustic realism, it simultaneously imposes substantial linguistic processing demands on non-native listeners. Voice cloning, by contrast, preserves speaker-specific identity cues such as timbre, rhythm, and turn-taking patterns while removing the need to decode a foreign language. This combination reduces extraneous cognitive load (RQ4), which in turn supports clearer comprehension (RQ1) and more favorable social perception (RQ2). Importantly, these effects interact rather than operate in isolation. Reduced mental effort appears to free attentional resources that listeners can reallocate toward tracking speakers, interpreting intent, and maintaining engagement, contributing to higher perceived intelligibility and social impression. This interaction helps explain the preference pattern observed in RQ3, where participants favored voice-cloned translation over original speech despite rating original speech as more natural. For non-native listeners, functional familiarity and cognitive efficiency outweighed absolute vocal authenticity. Participants frequently described cloned voices as "easier to follow," "less tiring," and "more coherent with the speaker," clarifying why preference diverged from raw naturalness ratings. These findings suggest that in cross-lingual communication, user experience is shaped not solely by naturalness but by how effectively a system supports identity continuity while minimizing processing demands. Voice cloning appears to occupy a favorable middle ground: less authentic than original speech, but more socially and cognitively usable than generic synthetic voices. This synthesis underscores the importance of holistically evaluating translation technologies, considering comprehension, perception, preference, and cognitive load as interdependent dimensions rather than independent outcomes.

5.2 Theoretical and Practical Implications

Our findings challenge the conventional approach to cross-lingual communication systems, which has prioritized semantic accuracy at the expense of vocal identity. The consistent advantages of voice cloning across comprehension, perception, preference, and cognitive load measures suggest that *identity preservation* can be considered a fundamental requirement rather than an optional enhancement in translation systems.

The reduced mental workload associated with voice cloning has important practical implications for extended meetings and complex discussions. As organizations increasingly rely on global virtual collaboration, technologies that minimize cognitive fatigue can

directly impact meeting effectiveness and participant engagement. The 40% reduction in mental effort we observed could translate to meaningful improvements in sustained attention and information retention during lengthy cross-lingual sessions.

The dissociation between naturalness (where original speech excelled) and social impression (where cloning outperformed generic TTS) suggests these are distinct perceptual dimensions that system designers can optimize independently. This finding is valuable for practical implementations, as it indicates that voice cloning can provide social benefits even while continuing to evolve toward naturalness parity with human speech.

5.3 Design Implications for Next-Generation Meeting Platforms

Identity-Preserving Translation as Default Mode: Given the advantages of voice cloning, meeting platforms can consider making identity-preserving translation the default option rather than a premium feature. The significant improvements in comprehension, reduced mental effort, and strong user preference justify this positioning.

Adaptive Voice Rendering: Systems can intelligently switch between voice modes based on context. For formal presentations where naturalness is paramount, original speech might be prioritized. For multi-party discussions where speaker tracking is crucial, voice cloning can be automatically activated to reduce cognitive load.

Enhanced Transparency Controls: While voice cloning offers clear benefits, platforms need to provide unambiguous indicators when synthesized speech is being used. Visual badges, audio watermarks, and detailed consent flows can be standard features to maintain trust and authenticity.

Cognitive Load Optimization: The mental workload reductions we observed suggest that voice cloning can be particularly emphasized in scenarios where cognitive demands are high — multi-speaker meetings, technical discussions, or extended sessions. Platforms could use meeting analytics to recommend voice modes based on predicted cognitive load.

Ethical Implementation Frameworks: The power of voice cloning necessitates robust ethical safeguards. Organizations can develop clear policies regarding voice enrollment, usage scope, and revocation procedures. Technical implementations can include watermarking, abuse detection, and usage auditing capabilities.

5.4 Limitations and Future Directions

Several limitations of the current study suggest important directions for future research. First, our evaluation used scripted meeting content in controlled conditions; real-world meetings with spontaneous conversation, overlapping speech, and background noise may yield different patterns. Future work should examine voice cloning performance in live, multi-party meetings with natural interaction dynamics. Second, while our sample represented technically sophisticated professionals, broader population sampling including varying language proficiencies, age groups, and cultural backgrounds would enhance generalizability. Particularly important would be inclusion of participants with different accessibility needs who might benefit disproportionately from identity-preserving translation. Third, the current study evaluated a specific voice cloning implementation. Comparative evaluations across different technical approaches (speaker adaptation, zero-shot cloning, few-shot learning) would help identify which architectural choices most impact user experience and cognitive load. Fourth, while we focused on immediate perceptual measures, longitudinal studies examining team performance, meeting outcomes, and technology adoption over time would provide valuable insights into real-world utility. Important questions remain about how voice cloning affects group dynamics, participation equality, and decision quality in cross-lingual teams. Additionally, while voice cloning preserved speaker identity cues, it did not fully convey emotional dynamics. Contemporary TTS systems often generate affect heuristically rather than reproducing authentic emotional states, which may explain why original speech retained advantages in perceived naturalness. Future work should examine emotion-aware speech synthesis and cross-lingual emotion transfer as complementary advances. Finally, ethical considerations demand continued attention. Future work should explore effective watermarking techniques, abuse detection mechanisms, and organizational policies that balance innovation with responsible deployment. Cultural differences in voice perception and privacy expectations also merit cross-cultural investigation.

6 Conclusion

This research demonstrates that voice cloning technology represents a significant advancement for cross-lingual meeting systems, offering measurable benefits in mental workload reduction, user experience enhancement, and communication effectiveness. Our findings establish that identity-preserving translation provides cognitive and social advantages beyond what conventional TTS systems can deliver, while acknowledging that naturalness remains an area for continued improvement. The consistent preference for voice cloning across multiple metrics suggests that speaker identity preservation addresses fundamental needs in cross-lingual communication that have been largely overlooked in current systems. By maintaining the personal characteristics that make human communication rich and engaging, voice cloning technology has the potential to make multilingual collaboration more natural, efficient, and socially connected. However, these technical capabilities must be paired with thoughtful design and ethical safeguards. The design implications outlined provide a roadmap for implementing identity-preserving translation in ways that respect user autonomy, promote transparency, and prevent misuse. As voice cloning technology continues to evolve, maintaining this balance between innovation and responsibility will be crucial for realizing its full potential in enhancing global communication. Looking ahead, we envision a future where cross-lingual systems seamlessly preserve the personal and social dimensions of communication while removing language barriers — creating meeting experiences that are not only comprehensible but truly collaborative across linguistic and cultural boundaries.

Acknowledgments

This work was funded by the Research Council of Norway (326907), the Swedish Foundation for Strategic Research (FUS21-0067), and the HORIZON-CL4-2021-HUMAN-01 ELSA project.

7 GenAI Usage Disclosure

During the preparation of this work, the authors used ChatGPT-4 for language polishing and grammar checking in certain sections of the manuscript. All research design, data collection, analysis, interpretation, and substantive writing were conducted by the human authors. After using this tool, the authors reviewed and edited the content carefully and take full responsibility for the publication's content.

References

[1] 2025. Understanding voice naturalness. *Trends in Cognitive Sciences* 29, 5 (2025), 467–480.

[2] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural Voice Cloning with a Few Samples. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc.

[3] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation.

[4] Dennis Becker, Lukas Braach, Lennart Clasmeier, Teresa Kaufmann, Oskar Ong, Kyra Ahrens, Connor Gäde, Erik Strahl, Di Fu, and Stefan Wermter. 2025. Influence of Robots' Voice Naturalness on Trust and Compliance. *J. Hum.-Robot Interact.* 14, 2, Article 29 (Jan. 2025), 25 pages.

[5] Nancy H. Brinson, Steven Holiday, and Jessica L. George. 2024. Response to Advertising Delivered by Voice Assistants: The Mediating Role of Persuasion Knowledge, Perceived Control, Social Presence, and Privacy Concerns. *Journal of Interactive Advertising* 24, 4 (2024), 344–367.

[6] Leo Chadburn. 2023. *Captions, characters, self-portraits: compositional approaches to the disembodied speaking voice and the voice-text-music relationship.* Ph. D. Dissertation. City, University of London.

[7] Sanyuan Chen, Shujie LIU, Long Zhou, Eric Liu, Xu Tan, Jinyu Li, sheng zhao, Yao Qian, and Furu Wei. 2025. VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers. https://openreview.net/forum?id=0bcRCD7YUx

[8] Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370* (2024).

[9] Xinjie Chen, Kai Fan, Wei Luo, Linlin Zhang, Libo Zhao, Xinggao Liu, and Zhongqiang Huang. 2024. Divergence-guided simultaneous speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. AAAI Press, 17799–17807.

[10] Ge Gao, Bin Xu, Dan Cosley, and Susan R. Fussell. 2014. How beliefs about the presence of machine translation impact multilingual collaborations. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, 1549–1560.

[11] Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, 852–863.

[12] Ge Gao, Naomi Yamashita, Ari M.J. Hautasaari, and Susan R. Fussell. 2015. Improving Multilingual Collaboration by Displaying How Non-native Speakers Use Automated Transcripts and Bilingual Dictionaries. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. 3463–3472.

[13] Genesis Gregorious Genelza. 2024. A systematic literature review on AI voice cloning generator: A game-changer or a threat? *Journal of Emerging Technologies* 4, 2 (2024), 54–61.

[14] Morton Ann Gernsbacher. 2015. Video captions benefit everyone. *Policy insights from the behavioral and brain sciences* 2, 1 (2015), 195–202.

[15] Marion Hersh, Barbara Leporini, and Marina Buzzi. 2024. A comparative study of disabled people's experiences with the video conferencing tools Zoom, MS Teams, Google Meet and Skype. *Behaviour & Information Technology* 43, 15 (2024), 3777–3796.

[16] Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International conference on machine learning*. PMLR, 10120–10134.

[17] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* 31 (2018).

[18] Kamil Deja and Ariadna Sanchez and Julian Roth and Marius Cotescu. 2022. Automatic Evaluation of Speaker Similarity. In *Interspeech 2022*. International Speech Communication Association (ISCA), 2348–2352.

[19] Victor Kenji, Anthony J. Lee, Daria Altenburg, David R. Feinberg, and Benedict C. Jones. 2022. The Role of Valence, Dominance, and Pitch in Perceptions of Artificial Intelligence (AI) Conversational Agents' Voices. *Scientific Reports* 12, 1 (2022), 22479.

[20] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. 2020. The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics* 14 (2020).

[21] Raja S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2013. Captions versus transcripts for online video content. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. Association for Computing Machinery, 1–4.

[22] Eun-Ju Lee. 2003. Effects of "gender" of the computer on informational social influence: the moderating role of task type. *International Journal of Human-Computer Studies* 58, 4 (2003), 347–362.

[23] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can computer-generated speech have gender? an experimental test of gender stereotype. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI EA '00)*. Association for Computing Machinery, New York, NY, USA, 289–290.

[24] Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. 2019. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence* 3, 4 (2019), 297–312.

[25] Marco Matassoni, Seraphina Fong, and Alessio Brutti. 2024. Speaker Anonymization: Disentangling Speaker Features from Pre-Trained Speech Embeddings for Voice Conversion. *Applied Sciences* 14, 9 (2024).

[26] Liz McCarron. 2021. Creating accessible videos: Captions and transcripts. *Communications of the Association for Information Systems* 48, 1 (2021), 19.

[27] Oksana Novytska, Hlib Romanchuk, Oleksii Vorobets, Uliana Zhornokui, Liubov Slyvka, and Valerii Bohdan. 2025. Translation of Subtitles: Neurolinguistic and Cognitive Aspects. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience* 16, 1 (2025), 229–242.

[28] Gary M Olson, Judith S Olson, Mark R Carter, and Marianne Storrosten. 1992. Small group design meetings: An analysis of collaboration. *Human–Computer Interaction* 7, 4 (1992), 347–374.

[29] Laura Orynbay, Bibigul Razakhova, Peter Peer, Blaž Meden, and Žiga Emeršič. 2024. Recent advances in synthesis and interaction of speech, text, and vision. *Electronics* 13, 9 (2024), 1726.

[30] Sara Papi, Matteo Negri, and Marco Turchi. 2023. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 13340–13356.

[31] Abhijeet Kumar Patel, Hardik Madnani, Sambhav Tripathi, Purushottam Sharma, and Vinod Kumar Shukla. 2024. Real-Time Voice Cloning: Artificial Intelligence to Clone and Generate Human Voice. In *International Conference on Information Technology*. Springer, 349–364.

[32] Sabid Bin Habib Pias, Ran Huang, Donald S. Williamson, Minjeong Kim, and Apu Kapadia. 2024. The Impact of Perceived Tone, Age, and Gender on Voice Assistant Persuasiveness in the Context of Product Recommendations. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 20, 15 pages.

[33] Victor Rosi, Emma Soopramanien, and Carolyn McGettigan. 2025. Perception and social evaluation of cloned and recorded voices: Effects of familiarity and self-relevance. *Computers in Human Behavior: Artificial Humans* 4 (2025), 100143.

[34] Mohammad Sarim, Saim Shakeel, Laeeba Javed, Mohammad Nadeem, et al. 2025. Direct Speech to Speech Translation: A Review. *arXiv preprint arXiv:2503.04799* (2025).

[35] Dave Sayers, Rui Sousa-Silva, Sviatlana Höhn, Lule Ahmedi, Kais Allkivi-Metsoja, Dimitra Anastasiou, Štefan Beňuš, Lynne Bowker, Eliot Bytyçi, Alejandro Catala, et al. 2021. The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies. (2021).

[36] Scott Schanke, Gordon Burtch, and Gautam Ray. 2024. Digital lyrebirds: Experimental evidence that voice-based deep fakes influence trust. *Management Science* (2024).

[37] Matthew Seita, Sooyeon Lee, Sarah Andrew, Kristen Shinohara, and Matt Huenerfauth. 2022. Remotely Co-Designing Features for Communication Applications using Automatic Captioning with Deaf and Hearing Pairs. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 460, 13 pages.

[38] Clay Spinuzzi. 2012. Working alone together: Coworking as emergent collaborative activity. *Journal of business and technical communication* 26, 4 (2012), 399–441.

[39] Giselle Spiteri Miggiani. 2024. Quality assessment tools for studio and AI-generated dubs and voice-overs. (2024).

[40] John C Tang, Chen Zhao, Xiang Cao, and Kori Inkpen. 2011. Your time zone or mine? A study of globally time zone-shifted collaboration. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. Association for Computing Machinery, 235–244.

[41] Sunday David Ubur. 2025. Augmenting Captions with Emotional Cues: An AR Interface for Real-Time Accessible Communication. arXiv:2504.17171 [cs.HC] https://arxiv.org/abs/2504.17171

[42] Vimal Kumar Vishwakarma. 2023. Translating cultural nuances: Challenges and strategies. *ELT Voices* 13, 2 (2023), 8268531.

[43] Haldun Vural. 2025. TRANSLATION-FOCUSED TECHNOLOGICAL COMPETENCE: TRADITION AND INNOVATION. *Cumhuriyet Üniversitesi Fen-Edebiyat Fakültesi Sosyal Bilimler Dergisi* 49, 1 (2025), 85–95.

[44] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023).

[45] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. I'd blush if I could: closing gender divides in digital skills through education. (2019), 1–145.

[46] Kevin JP Woods, Max H Siegel, James Traer, and Josh H McDermott. 2017. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics* 79, 7 (2017), 2064–2072.

[47] Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*. Association for Computing Machinery, New York, NY, USA, 515–524.

[48] Taojie Yin. 2025. Has the use of AI-translated live captions in simultaneous interpreting changed the role of the interpreter? A study based on professional interpreters' perceptions. *The Translator* 31, 2 (2025), 214–231.

[49] Yongle Zhang, Dennis Asamoah Owusu, Emily Gong, Shaan Chopra, Marine Carpuat, and Ge Gao. 2021. Leveraging Machine Translation to Support Distributed Teamwork Between Language-Based Subgroups: The Effects of Automated Keyword Tagging. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, Article 381, 6 pages.

[50] Ziqiang Zhang, Long Zhou, Chengyi Wang, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling.