

Perceptual comparison of emotional Korean speech: Human vs. AI-generated voices

Na-Young Ryu (Penn State U; nayoung.ryu@psu.edu)
Suyeon Yun (Chungnam National U; suyeon.yun@cnu.ac.kr)

Introduction

- Recent TTS systems achieve near-human clarity but still struggle with emotional expressiveness – especially in underrepresented languages like Korean.
- Little is known about how Korean listeners actually perceive these AI emotions. This study fills that gap.
- This study examines how native Korean listeners distinguish human from AI voices and evaluate their naturalness, emotional accuracy, and contextual appropriateness.

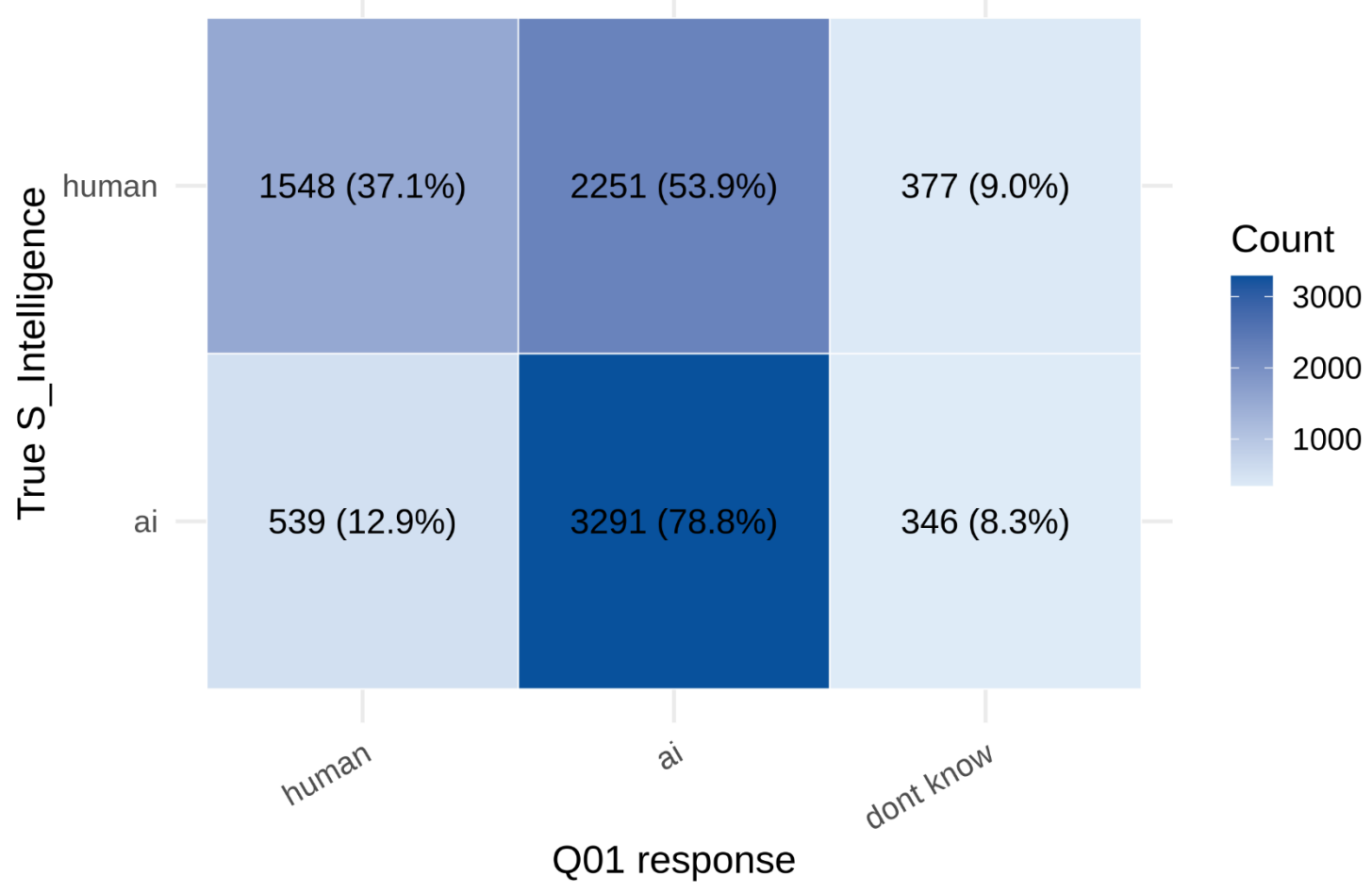
Methodology

- Participants:** 87 native Korean speakers (63f, 24m; mean age=21.5)
- Stimuli:** 96 Korean sentences
 - 2 sentences read with emotion and in a neutral tone
 - 6 emotions (happiness, sadness, fear, anger, surprise, disgust)**
 - 2 human** (1f, 1m; professional voice actors) + **2 AI speakers** (1f, 1m)
 - AI voice synthesized using **Zyphra’s Zonos-v0.1**
 - Open-source accessibility
 - Allows direct control over emotion parameters
 - High-fidelity voice cloning capability
- Procedure:** Online experiment
 - Auditory stimuli were presented in a randomized order
 - After listening to each stimulus, participants answered 8 questions:
 - Whose voice do you think this is? Human or AI?
 - How natural did this voice sound? (1-9)
 - If the voice sounded unnatural, which aspects contributed to that perception? – emotional expression, intonation and rhythm, pronunciation accuracy, voice quality, speaking rate
 - How accurate was the pronunciation? (1-9)
 - How natural was the intonation? (1-9)
 - What emotion did you perceive most strongly in this voice? – happiness, sadness, anger, surprise, fear, disgust, neutral
 - Did the expressed emotion match the content of the sentence? (1-9)
 - How would you evaluate the emotional expression in this voice? – exaggerated, natural, insufficient

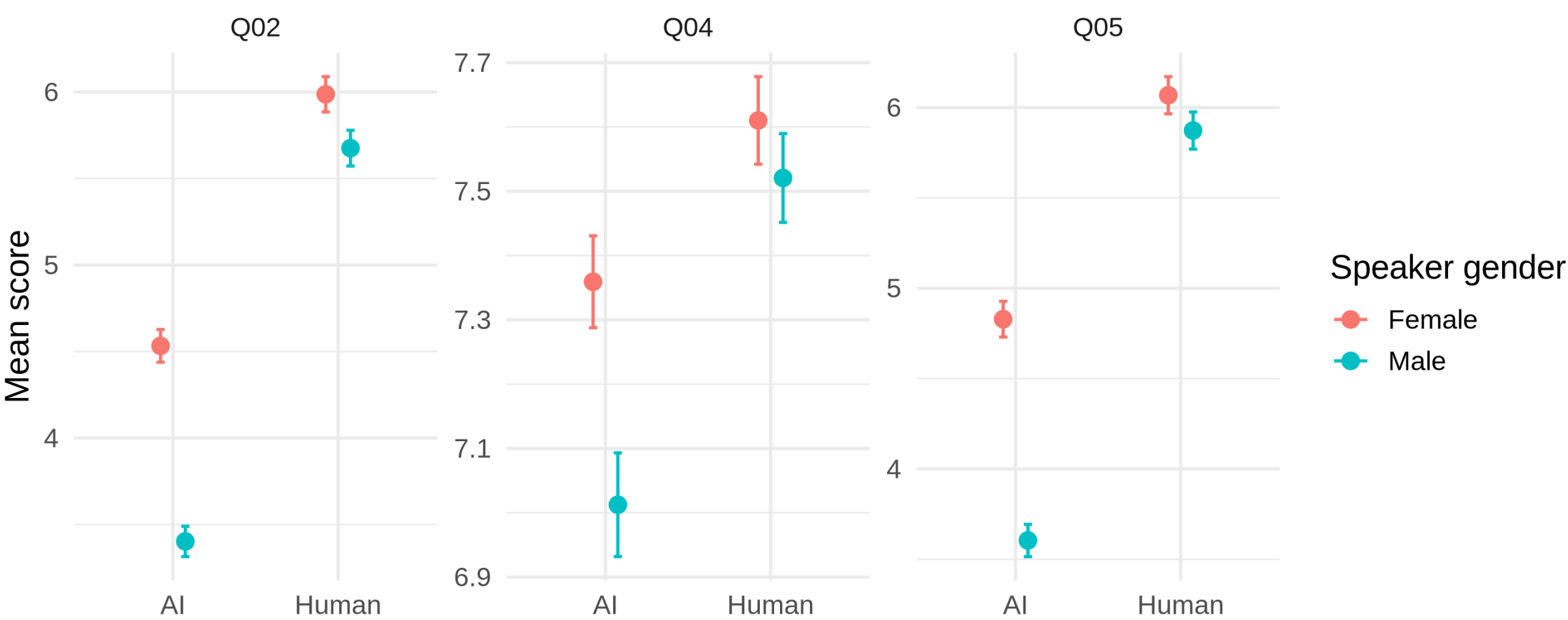


Results

- Q01 (**Human vs. AI identification**): High accuracy in identifying AI voices as AI, but significantly lower accuracy in identifying human voices as human → AI voice synthesis technology has reached an advanced level
 - No significant effect of speaker gender and frequency of AI voice use

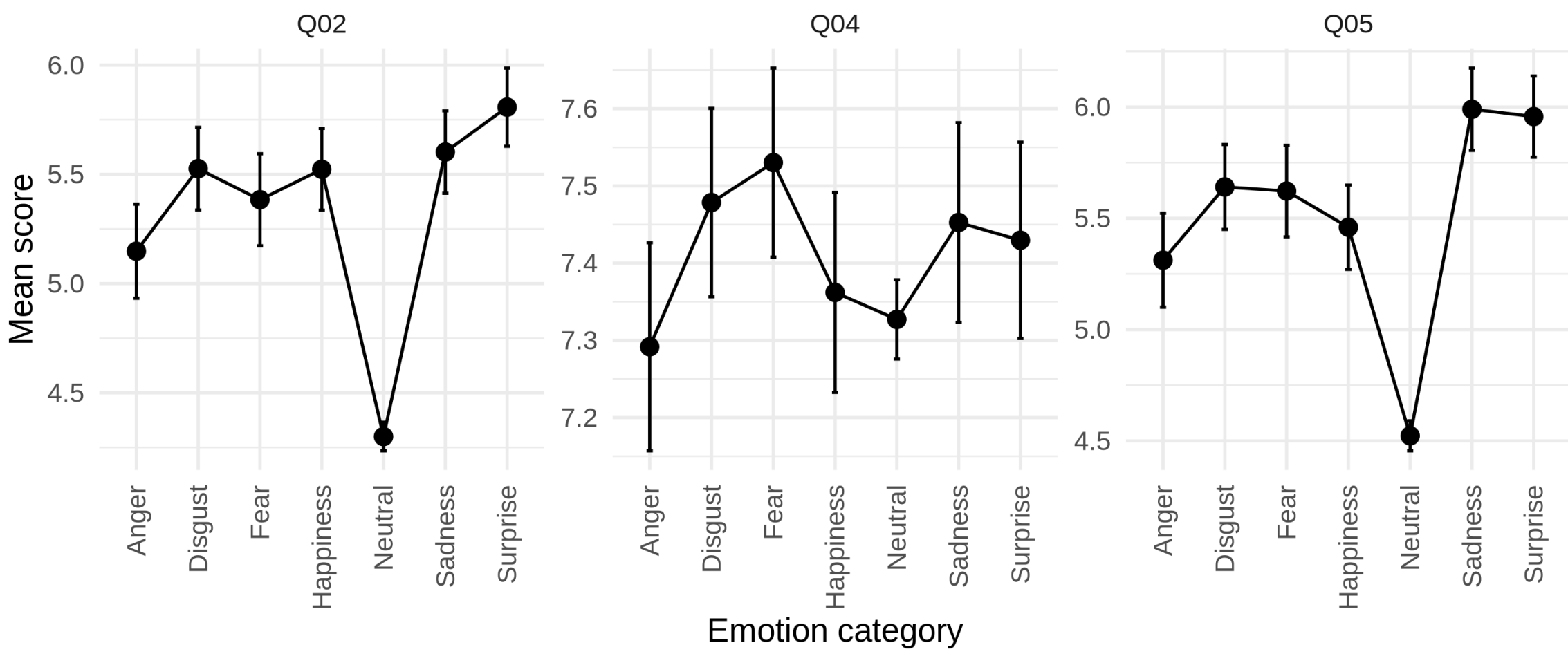


- Q02 (**overall naturalness**)
 - Q04 (**pronunciation accuracy**)
 - Q05 (**intonational naturalness**)
- Human > AI (9-point Likert scale)
- Male voices were rated lower than female, especially in AI voices

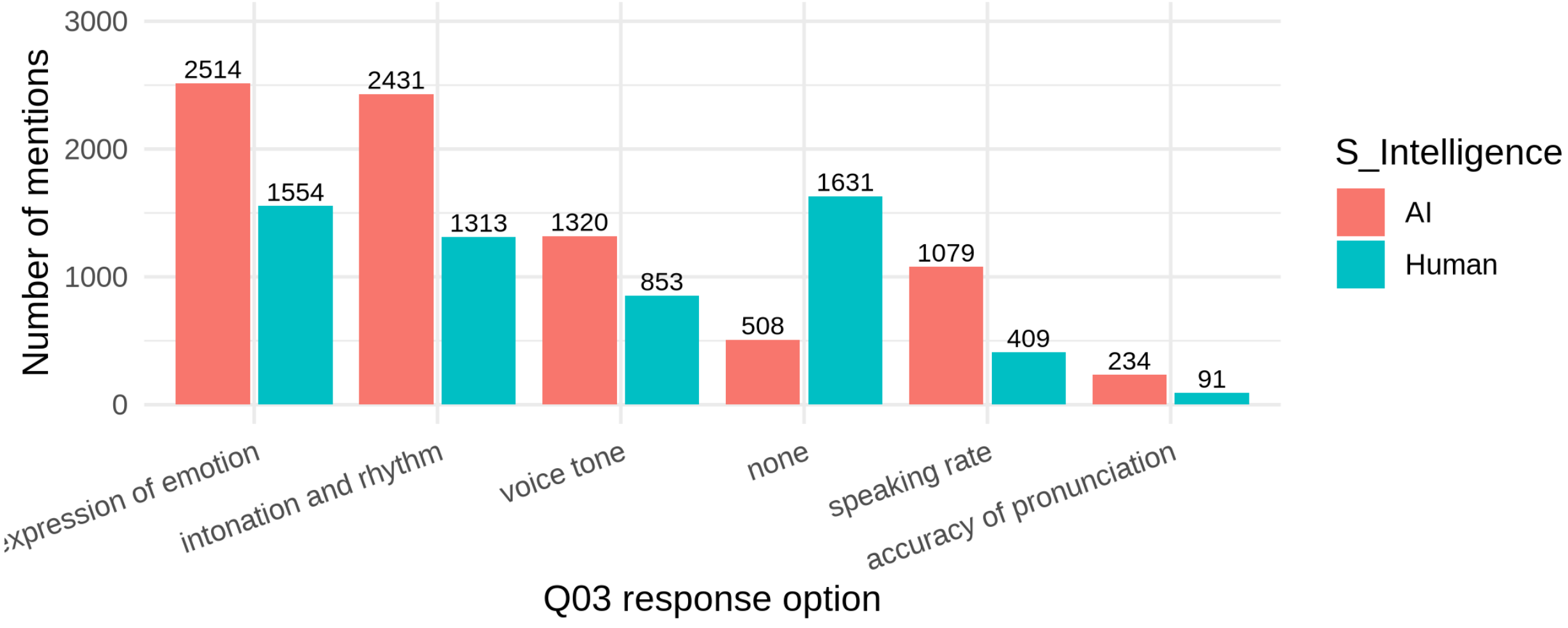


AI-generated Korean speech still struggles significantly to convey specific emotions authentically.

- “Neutral” was rated lower for Q02 (overall naturalness) and Q05 (intonational naturalness)
- Other emotions don’t show clear differences



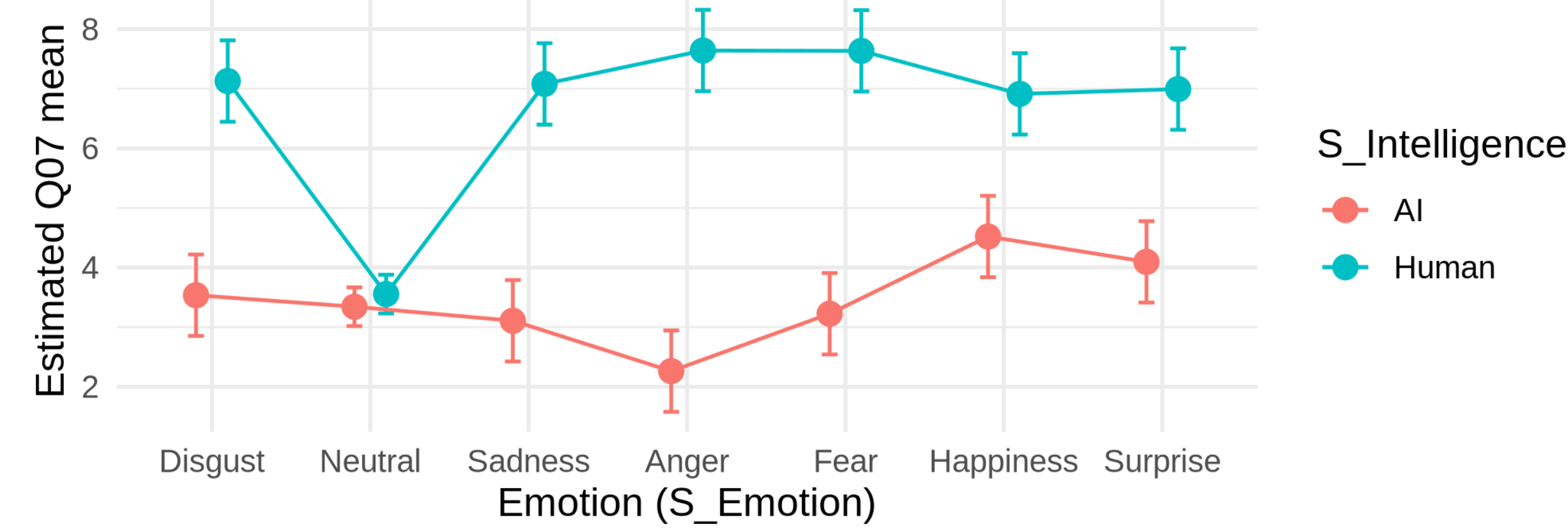
- Q03 (**Reasons for perceived unnaturalness**)
 - ‘Emotional expression’** and **‘Intonation and rhythm’** are the most critical factors contributing to perceived unnaturalness in AI voices



- Q06 (**Emotion recognition accuracy**): 48.6% overall
 - Human 65.54% vs. AI 31.66% → significantly higher for human voices for every emotion**
 - Only in “happiness”, male voices are recognized more accurately than female voices; no gender difference for other emotions
 - For **AI**: Neutral, Happiness > Disgust, Fear, Sadness > Anger, Surprise
 - For **Human**: Anger, Fear, Sadness > Happiness, Surprise > Neutral, Disgust



- Q07 (**Emotion-content appropriateness**) (9-point scale)
 - Human 5.39 vs. AI 3.40 → significantly higher for human voices**
 - Human > AI for all emotions except Neutral
 - For **AI**: Happiness, Surprise > Anger, Neutral
 - For **Human**: All non-neutral emotions > Neutral



- Q08 (**Emotional expressiveness**):
 - For **AI**: **Insufficient 62.7%** > natural 15.6% > excessive 7.3%
 - For **Human**: Natural 43% > insufficient 39.7% > excessive 8.1%

Discussion

- Human vs. AI distinction: Native Korean listeners reliably identify AI voices, yet often confuse human voices with AI, indicating current TTS sophistication but incomplete human-likeness
- Quality gaps: Human voices significantly outperform AI across overall naturalness, pronunciation accuracy, and intonational quality – especially in emotional expression and intonation
- Emotion recognition: The largest performance gap appears in emotion recognition accuracy and emotion-content appropriateness, where human voices outperform AI across most emotions
- Prosody challenges: Emotional expressiveness remains the major bottleneck for Korean TTS, highlighting the importance of prosody for naturalistic emotional speech synthesis
- Technology direction: Results underscore the need for richer prosodic and affective modeling to advance AI voice quality



Scan to download a digital copy of this poster.

The 6th Joint Meeting of
the Acoustical Society of America
and the Acoustical Society of Japan
December 1-5, 2025