Differences in Gradient Emotion Perception: Human vs. Alexa Voices

Michelle Cohn, Eran Raveh, Kristin Predeck, Iona Gessinger, Bernd Möbius, Georgia Zellou
Overview

1. Background
2. Current Study
3. Analysis & Results
4. Discussion
5. Conclusion
Overview

1. Background
2. Current Study
3. Analysis & Results
4. Discussion
5. Conclusion
Interaction with voice-activated assistants (e.g., Alexa, Google Assistant, Siri) increased in recent years

→ but how similar are these interactions to human-human interactions?
Emotion and CASA

- Computers Are Social Actors (CASA) theoretical framework (Nass et al. 1997)
  - people hear ‘emotion’ in text-to-speech (TTS) voices (Nass et al., 2001; Brave et al. 2005, Bartneck 2001)
Emotion and CASA

- Computers Are Social Actors (CASA) theoretical framework (Nass et al. 1997)
  - e.g., people hear ‘emotion’ in TTS voices (Nass et al., 2001; Brave et al. 2005, Bartneck 2001)
  - voice-AI (e.g., Siri, Alexa) and emotion (e.g. Cohn et al., 2019)
Emotion and CASA

- Computers Are Social Actors (CASA) theoretical framework (Nass et al. 1997)
  - e.g., people hear ‘emotion’ in TTS voices (Nass et al., 2001; Brave et al. 2005, Bartneck 2001)
  - voice-AI (e.g., Siri, Alexa) and emotion (e.g. Cohn et al., 2019)

  - ... but voice-AI / humans → differences (Raveh et al., 2019; Cohn et al., 2019)
Variation in personification

- Vocal alignment toward voice-AI (like Apple Siri) systems depends on the users’ cognitive processing style (Snyder et al. 2019)

- Individuals interacting with the same robot receptionist communicated differently depending on their attitude towards the virtual interlocutor: as being more ‘human-social’ or a ‘computational-tool’ (Lee et al. 2010)

.... Variation for displays of emotion by a voice-AI system?
Overview

1. Background
2. Current Study
3. Results
4. Discussion
5. Conclusion
Research questions

**Q1:** Do listeners perceive acoustic variations conveying different levels of emotional state similarly for human and TTS voices?

**Q2:** Does an individual’s gradient perception of TTS voice emotion vary according to the degree to which they personify the system; that is, are listeners better at perceiving emotion for interlocutors they deem as being more ‘human-like’?
Current study

- Norming Study = select sentences
- Perception Study
Norming pre-study

- **Goal**: obtain ratings of emotionally neutral sentences
- 120 sentences from previous work
- Ratings from 48 English native speakers on a 0 (negative) to 100 (positive) scale
- Average scores are provided in the paper’s supplementary material

Examples of “true neutral” sentences (average score 50):

“*A bag is in the room.*”  
“The sailor swabbed the deck.”  
“They played a game of cat and mouse.”
Current study

- Norming Study
- Perception Study
Perception study — Stimuli & Participants

Stimuli
- 15 most neutral sentences (mean 49.9) produced by
  - a female human speaker
  - Alexa’s female US-English voice
- Re-synthesized at three equally-spaced happiness levels
- 15 sentences X 3 levels X 2 speakers = 90 stimuli in total

Participants
- 99 English native speakers
  - 71% female
  - Mean age 20 years old
  - 82% used voice-AI systems
Perception study — Procedure I

- Completed online (via Qualtrics)
- Audio calibration step
Perception study — Procedure I

Voice-AI anthropomorphism:
- Rating Alexa’s voice based on a non-manipulated utterance for machine-/human-like, artificial/natural, eerie/comforting, and cold/warm
Perception study — *Procedure II*

Comparison trials

- In each trial
  - the participant listened to either “Alexa” or “Amanda” (indicated by a silhouette)
  - a neutral sentences was played in one of the three happiness levels
  - The participant rated the valence and arousal of the utterance (scale 0-100)

- Each participant rated all 90 stimuli
Perception study — Procedure II

- Example sentences for different valence / arousal conditions:
  0 vs. 66%; human voice / Alexa
Overview

1. Background
2. Current Study
3. Analysis & Results
4. Discussion
5. Conclusion
Perception study — *Analysis*

- Modeled valence/arousal in separate linear mixed effects models (*lme4* R package)

\[ \text{\texttt{lmer(Va}l\texttt{ence} \sim \text{Happiness Level} \times \text{Interlocutor} + (1+\text{Interlocutor} | \text{Subject}))} \]
Perception study – Results

Perceived valence

Valence score (0-100)

Happiness Level

more positive

Alexa

Human
Perception study – Results

\[ \text{lmer(Valence} \sim \text{Happiness Level} \ast \text{Interlocutor} + (1 + \text{Interlocutor} \mid \text{Subject}) \) \]

- **Interlocutor** \((p < 0.01)\)
  - Less for Human than Alexa

- **Happiness Level** (all \(p < 0.001\))
  - Higher at 33% and 66% (rel. 0%)

- **Interlocutor \ast Happiness Level** (all \(p < 0.001\))
  - Larger increase for Human at 33% and 66%

(see Table 1 in the paper)
Perception study – Results

\[ \text{lmer(Arousal} \sim \text{Happiness Level} * \text{Interlocutor} + (1 + \text{Interlocutor} | \text{Subject})) \]

Interlocutor (N.S.)
- No diff. for Human & Alexa

Happiness Level (all \( p < 0.001 \))
- Higher at 33% and 66% (rel. 0%)

Interlocutor * Happiness Level (all \( p < 0.01 \))
- Larger increase for Human at 33% and 66%

(see Table 2 in the paper)
Perception study – Results

**Perceived valence**

- **Valence score (0-100)**
- More positive

**Perceived arousal**

- **Arousal score (0-100)**
- More excited

- **Happiness Level**

- **Alexa**
- **Human**
Anthropomorphism — Analysis

- Calculated a composite anthropomorphism score
  - machine-/human-like (100), artificial/natural (100), eerie/comforting (100), cold/warm (100)
- Analyzed the subset of data on Alexa
- \textit{lmer}(\text{Valence} \sim \text{Anthro} \times \text{Happiness Level} \times \text{Interlocutor} + (1 | \text{Subject}))
Anthropomorphism — Results

Happiness Level * Anthro Score
($p < 0.01$)

- More positive at 0% with increasing anthro. score
Anthropomorphism – Results

AROUSAL:
Happiness Level * Anthro Score
(p < 0.01)
- Less excitement at 66% with increasing anthro score
Overview

1. Background
2. Current Study
3. Results
4. Discussion
5. Conclusion
Discussion

- Listeners perceive emotion **gradiently** in **human** and **Alexa** voices
  - Both dimensions (valence + arousal): **human** voices
  - ... but limited to arousal for **Alexa** voices

→ supports ‘**Computers are Social Actors’ (CASA)** (Nass et al., 1994; 1997)
Discussion

- Why only a gradient effect for arousal for the Alexa voice?
  - Limit to technology personification?
  - Baseline valence of the Alexa voice? (fairly positive overall)

- ... but some limitations (serving as avenues for future research):
  - Only 2 voices (one in each category) → multiple human / TTS voices
  - Type & magnitude of emotion → look at other types of emotions, larger variations
  - Passive listening task → diff. In more dyadic, social interaction?
Discussion

- Some evidence for individual variation in emotion perception based on *anthropomorphism*
  
  - Greater personification:
    
    → rated *Alexa* voice as more ‘positive’ *before* manipulations (0%)
    
    → rated *Alexa* voice at 66% as *less* ‘excited’

- ... but not balanced by *gender*: may be a contributing factor
Overview

1. Background
2. Current Study
3. Results
4. Discussion
5. Conclusion
Conclusion

- Some dimensions of emotion perception for human voices also observed for Alexa voices

  .... but a nuanced relationship that requires further study
Acknowledgments

- Thank you to our participants and funding sources
  - National Science Foundation SBE Postdoctoral Research Fellowship (1911855) to MC.
  - Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID MO 597/6-2 and STE 2363/1-2.
Differences in Gradient Emotion Perception:
Human vs. Alexa Voices

Michelle Cohn, Eran Raveh, Kristin Predeck, Iona Gessinger, Bernd Möbius, Georgia Zellou
Differences in Gradient Emotion Perception: Human vs. Alexa Voices

Michelle Cohn, Eran Raveh, Kristin Predeck, Iona Gessinger, Bernd Möbius, Georgia Zellou
• Native English speakers ($n=99$)

• Heard semantically neutral sentences (*selected via a norming study*):
  ○ a human / Amazon Alexa

• **Manipulated in 3 levels of ‘happiness’**: 0% (no change), 33% and 66% ‘happier’
● Native English speakers ($n=99$)

● Heard semantically neutral sentences (*selected via a norming study*):
  ○ *a human / Amazon Alexa*

● **Manipulated in 3 levels of ‘happiness’**: 0% (no change), 33% and 66% ‘happier’
- Native English speakers \((n=99)\)
- Heard semantically neutral sentences (selected via a norming study):
  - a human / Amazon Alexa
- Manipulated in 3 levels of ‘happiness’: 0% (no change), 33% and 66% ‘happier’
Rating the voices

Please rate the voice on the following dimensions:

Machine-like
0 10 20 30 40 50 60 70 80
Human-like
90 100

Please rate how machine-/human-like the voice sounds (0=machine-like, 100 = human-like)

Cold
0 10 20 30 40 50 60 70 80 90 100
Warm

Please rate how cold/warm the voice sounds (0=cold, 100 = warm)

Artificial
0 10 20 30 40 50 60 70 80
Natural
90 100

Please rate how artificial/natural the voice sounds (0=artificial, 100 = natural)

Eerie
0 10 20 30 40 50 60 70 80
Comforting
90 100

Please rate how eerie/comforting the voice sounds (0=eerie, 100 = comforting)