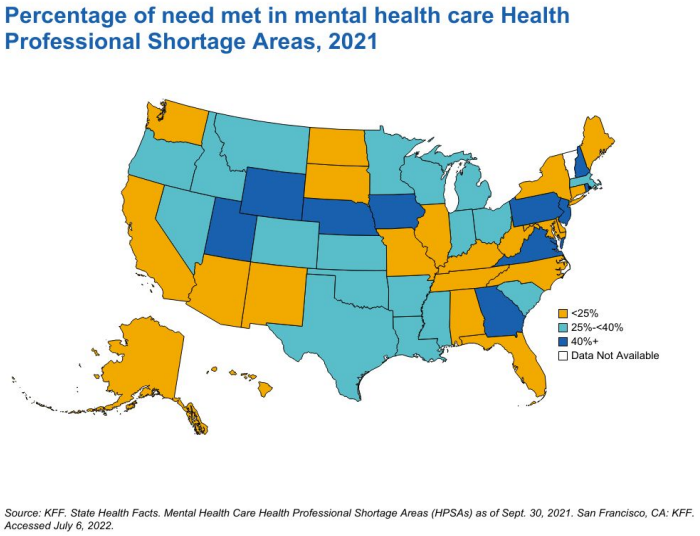
Learning to Generate Context-Sensitive Backchannel Smiles for Embodied AI Agents with Applications in Mental Health Dialogues

Maneesh Bilalpur, Mert Inan, Dorsa Zeinali,
Jeffrey Cohn and Malihe Alikhani

Machine Learning for Cognitive and Mental Health Workshop,
AAAI 2024

AI addressing Mental Health needs

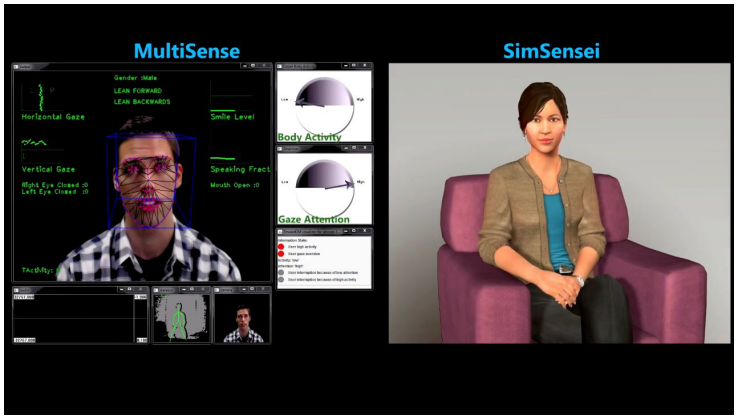
- Only **30% of Americans** in need have access to mental health care.
- AI has made great progress in symptom detection and monitoring treatment efficacy.



Modi et al. (2022)

Backchannels and Context

- Achieving common ground is important for the success of an interaction.
- Backchannel smiles help achieve the common ground.
- Not all backchannel smiles are created equally. Context defines the right fit.



DeVault et al. (2014)

What are Backchannels?

- Backchannels (BC) are listener behaviors.
- They express engagement, agreement and emotional response.
- **Rapport builders:** too short or too long might lead to conversational failures.
- Speaker prosody-based **rules** for BC generation. **Mimicking** speaker behavior or **discriminative** approach for production.

Research Questions

- Hypothesis: Do **speaker and listener behaviors influence** backchannel smiles?
- Can **generative models leverage salient behaviors** and improve the performance?

RealTalk Dataset

Lack of open-source datasets with patient-therapist interactions.

- YouTube-based video dataset of intimate dyadic interactions.

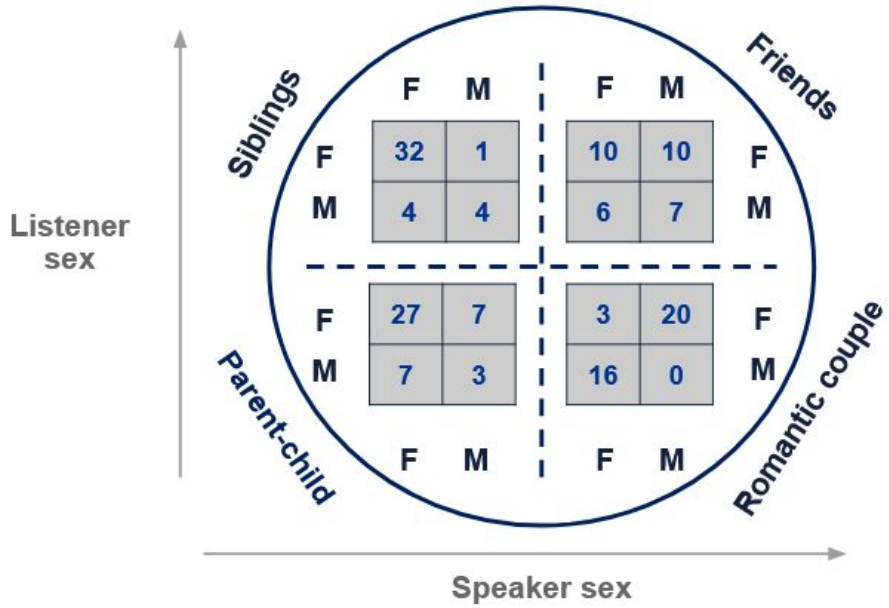
- Questions about:
  - Family relations
  - Dreams
  - Mental health etc.

Question: When do you feel closest to me?

Snippet from RealTalk (Geng et al. 2023) curated from the *SkinDeep* YouTube channel.

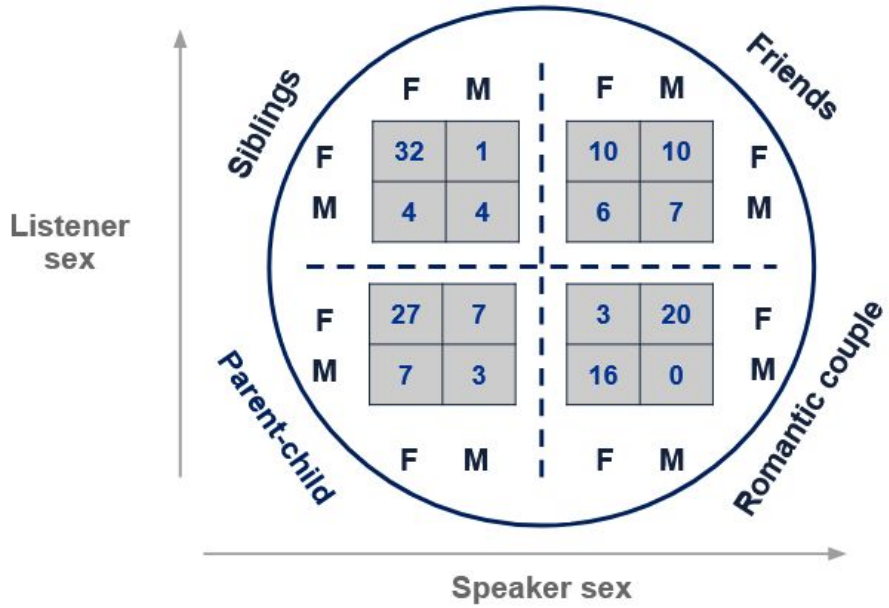
Annotating RealTalk: Identifying Backchannel Smiles

- **191 backchannel smiles from 48 interactions.**
- **83 % smiles had A-level or higher intensity.**



Annotating RealTalk: Identifying Backchannel Smiles

- **191 backchannel smiles from 48 interactions.**
- **83 % smiles had A-level or higher intensity.**



- How are smiles affected by dyadic characteristics like sex and relationship type?
- How do context-cues affect them?
- Can we leverage them in a generative approach?

Sex and Relationships Affect Backchannel Smiles

- We considered sex of the individuals and the nature of their interpersonal relationship for their effect on smile intensity and duration.
  - **Duration** of smiles differ by **listener sex** and **listener sex * relationship**
  - **Male listeners with their sibling (regardless of the sex) express longer BC smiles** ($p<0.05$).
  - **Intensity** marginally differs by speaker sex: Male speakers evoke less intense smiles than female speakers.

How Context Cues Affect Backchannel Smiles?

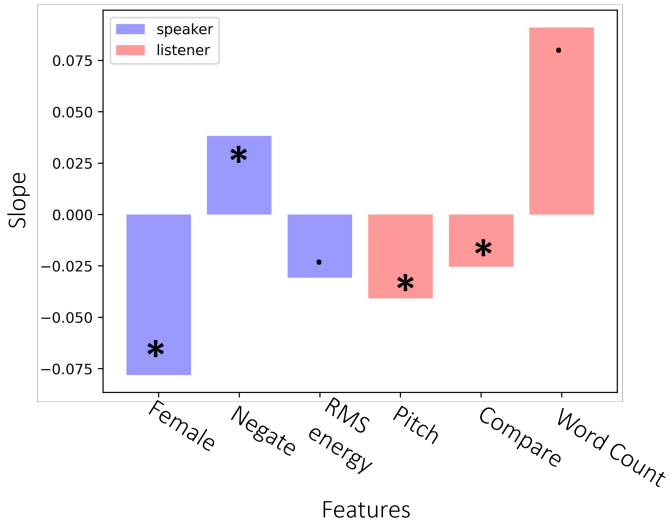
- Both speaker and listener preceding turns effect BC smiles.
- Prosodic and language cues from the turns were used to identify significant predictors.

Prosody	Language
Mean vocal pitch	Word-count
Range of pitch	Negations, Comparisons and Interrogatives
Loudness	Valence
	Focus on past, present and future

Effect of Context Cues on Backchannel Smiles

- When a speaker used **negations** the BC smiles were bigger. **Women speakers** evoked smaller BC smiles.

- When listeners used **high pitch voice** or **comparison words**, the BC smiles were smaller. If the listeners were **talkative**, the BC smiles were bigger.



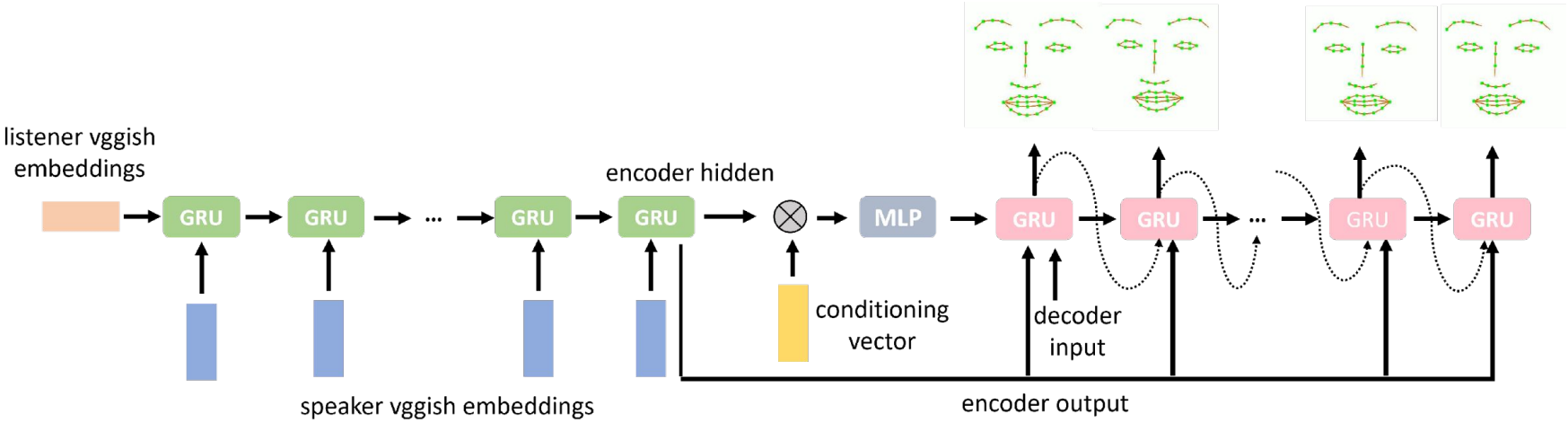
R^2=0.243.
"*" denotes $p<0.05$
"." denotes $p<0.1$.

Duration had no significant outcomes.

Context-Sensitive Backchannel Smile Generation

- Can generative models produce backchannel smiles that are context sensitive?
- Can we improve generative models from our understanding of context and backchannel smiles?

Proposed Architecture



Input:

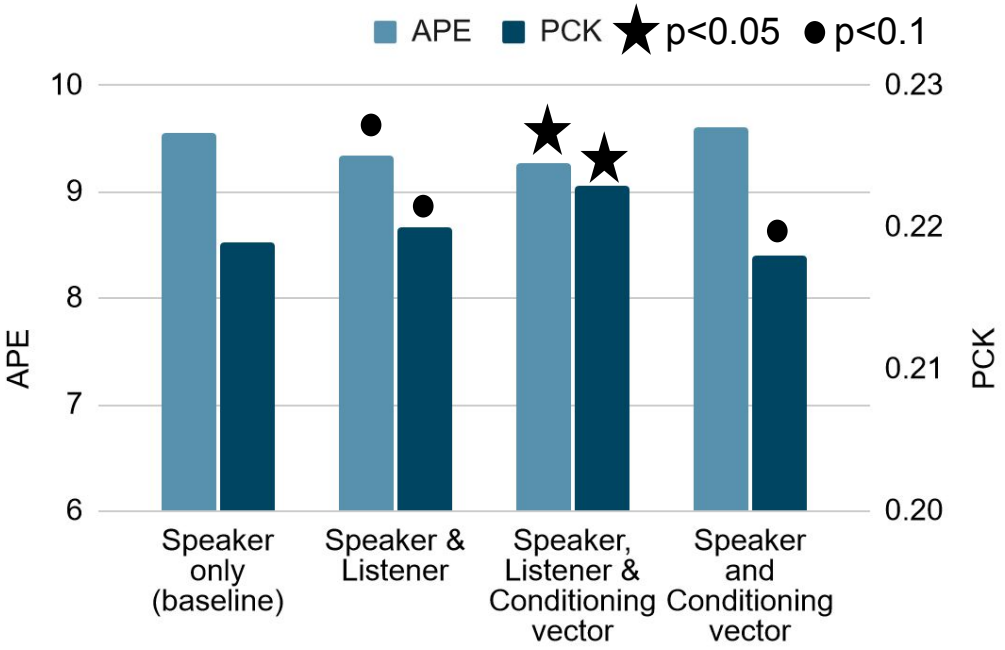
- **vggish embeddings:** turn-specific audio embeddings from vggish model for speaker and listener.
- **Conditioning vector:** speaker features (sex, negations), listener features (comparisons, mean pitch, word-count).

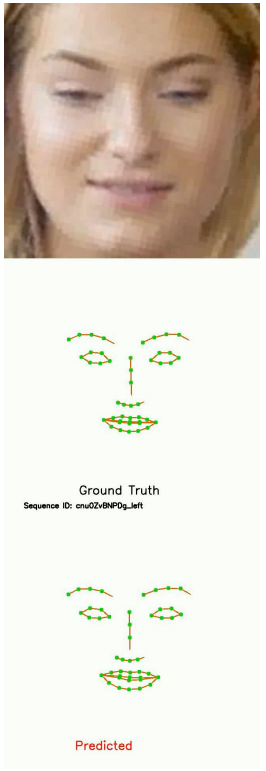
Output:

- Autoregressive prediction of 49 facial landmarks optimised with MSE loss.

Automatic Evaluation

- **Metrics:** Average Pose Error (APE) and Proximally Correct Keypoints (PCK).

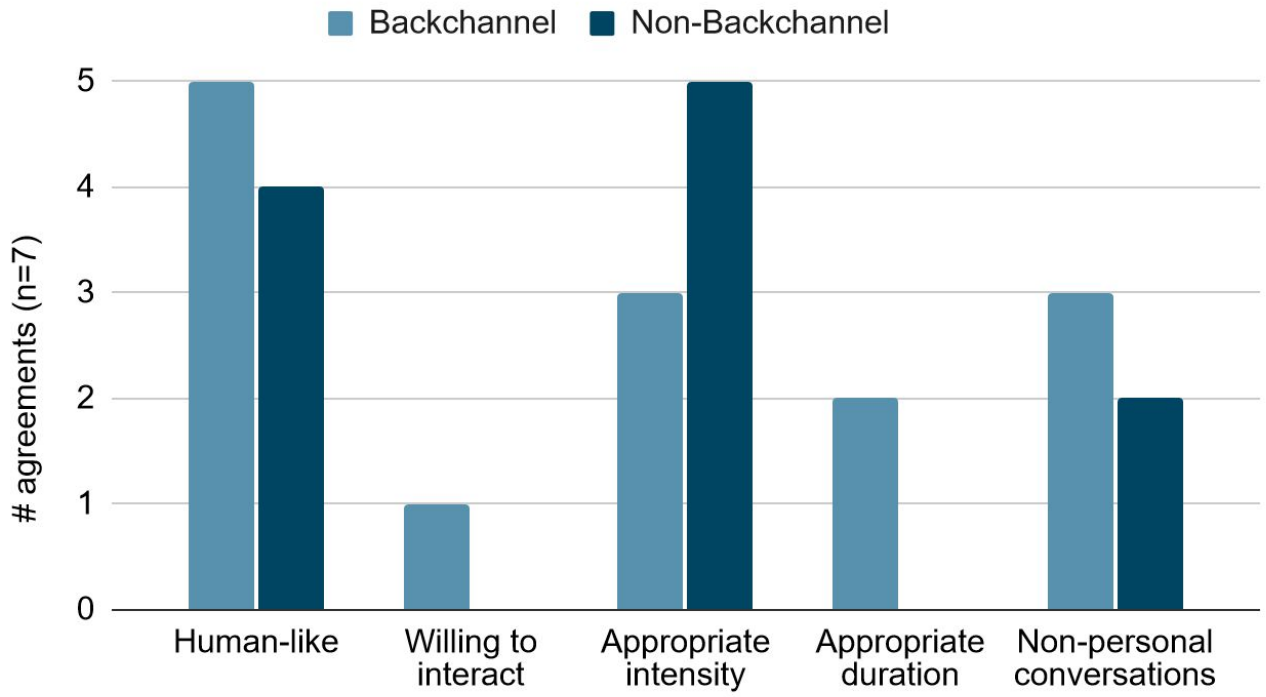




- Generations of **low-intensity smiles are better.**

Human Evaluation with an Embodied Agent

- Watch a Furhat agent interacting with a *human **with backchannels** and **without*** smiles.



Discussion

- Speaker-listener prosodic and linguistic behaviors, and demographics are significant predictors of backchannel (BC) smiles.
- BC smile generation had significant improvements when predictors from speaker and listener were used with their audio embeddings.
- BC smiles that co-occur with vocal activity are harder to predict.
- **Limitations:** improving annotation reliability, one smile-per-person assumption, tracking challenges, advanced generation models.

Contributions

- We annotated video dataset of diverse dyads for backchannel (BC) smiles.
- Our statistical analyses identified the affect of sex, relationship, and context-cues.
- We found that leveraging select context-cues generate better BC smiles.
- We bridged the gap between generation and realization by transferring facial landmarks to an embodied agent.
- We found that humans preferred agents with BC smile behaviors for non-personal conversations.

Thank you

Authors:

Acknowledgement: NIH award MH R01-096951.

Code and Dataset

SCAN ME