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

Maneesh Bilalpur[1,*], Mert Inan[2], Dorsa Zeinali[2], Jeffrey F. Cohn[1] and Malihe Alikhani[2]

[1]University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[2]Northeastern University, Boston, Massachusetts, USA

Abstract

Addressing the critical shortage of mental health resources for effective screening, diagnosis, and treatment remains a significant challenge. This scarcity underscores the need for innovative solutions, particularly in enhancing the accessibility and efficacy of therapeutic support. Embodied agents with advanced interactive capabilities emerge as a promising and cost-effective supplement to traditional caregiving methods. Crucial to these agents' effectiveness is their ability to simulate non-verbal behaviors, like backchannels, that are pivotal in establishing rapport and understanding in therapeutic contexts but remain under-explored. To improve the rapport-building capabilities of embodied agents we annotated backchannel smiles in videos of intimate face-to-face conversations over topics such as mental health, illness, and relationships. We hypothesized that both speaker and listener behaviors affect the duration and intensity of backchannel smiles. Using cues from speech prosody and language along with the demographics of the speaker and listener, we found them to contain significant predictors of the intensity of backchannel smiles. Based on our findings, we introduce backchannel smile production in embodied agents as a generation problem. Our attention-based generative model suggests that listener information offers performance improvements over the baseline speaker-centric generation approach. Conditioned generation using the significant predictors of smile intensity provides statistically significant improvements in empirical measures of generation quality. Our user study by transferring generated smiles to an embodied agent suggests that agent with backchannel smiles is perceived to be more human-like and is an attractive alternative for non-personal conversations over agent without backchannel smiles.

1. Introduction

Fewer than a third of the US population has sufficient access to mental health professionals [1]. This highlights the need for additional resources to help mental health professionals meet the community's demands. Problems like symptom detection and evaluating treatment efficacy have made great strides with AI [2, 3, 4] and the mental health community can greatly benefit from this AI intervention. Embodied agent-based systems due to their multimodal behavioral capabilities are a promising solution to support such mental health needs. However, the development of such systems presents numerous challenges. These include the scarcity of mental health-related datasets, limited access to domain experts for designing reliable and robust systems, and the ethical considerations crucial to their design and adaptation. Among such challenges, one aspect that stands out is the agent's ability to establish a common ground with users. Addressing this is particularly crucial when the agent functions as a listener. Effective grounding in such scenarios relies heavily on multimodal non-verbal behaviors like backchannels. These subtle yet impactful cues are pivotal in building rapport and understanding between the user and the agent. Hence, understanding and incorporating these behaviors into embodied agents is not only challenging but also essential for creating a

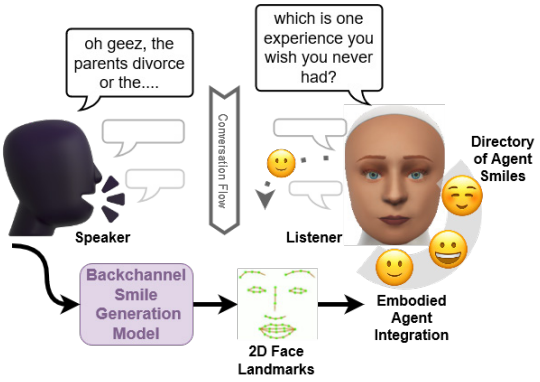

Figure 1: Overview of steps for backchannel smile generation in an embodied agent in a human-agent interaction: Speaker and listener (agent) turns are used to generate the listener's response facial expression as landmarks. The landmarks are then integrated with the embodied agent and added to the conversation flow represented as a dotted arrow.

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada.
*Corresponding author.
mab623@pitt.edu (M. Bilalpur); inan.m@northeastern.edu (M. Inan); zeinali.d@northeastern.edu (D. Zeinali); jeffcohn@pitt.edu (J. F. Cohn); m.alikhani@northeastern.edu (M. Alikhani)

supportive and empathetic environment for individuals seeking mental health support. Addressing these challenges can pave the way for more effective, accessible, and empathetic digital mental health interventions.

In dyadic conversations, at any given time one person may have the floor (i.e., is speaking) while the other is listening. Backchannels (BC) refer to behaviors of the listener that do not interrupt the speaker. BCs signal attention, agreement, and emotional response to what is said. Inappropriate BC smiles such as ones that appear too short or too long or for which the timing appears "off" can disrupt the conversational rapport and result in unsuccessful or disrupted conversations. Our objective is to understand appropriate BC smiles from dyadic conversations and how an embodied agent can employ them when interacting with a human.

Conversational agents typically realize BC smiles using rule-based systems, discriminative approaches, or sometimes simply mimicking the smiles of the speaker. Mimicking, however, fails to generalize to situations that require a contextually relevant smile. And rule-based and discriminative approaches offer limited coverage due to the diversity of smiles [5].

We present a generative approach for BC smiles in listeners to address these limitations and enable contextually relevant BC smiles in embodied agents. An overview of the approach is presented in Figure 1. Unlike existing works that solely depend on speaker behavior for BC production (see related work section), we use both speaker and listener behaviors to study how they affect the intensity and duration of the BC smile. We use cues from prosody, language, and the demographics of dyads to identify statistically significant predictors (referred to as a conditioning vector) of smiles. In addition to the audio features from both interaction participants, we leverage the conditioning vector in generating the BC smiles. In this paper, we:

1. Annotate backchannel smiles in a face-to-face interaction dataset[1] of dyads that differ in their composition of biological sex and type of relationship.
2. Present our statistical analysis to identify various speaker and listener-specific cues that significantly predict the duration and intensity of backchannel smiles.
3. Generate backchannel smiles using an attention-based generative model that uses the listener and speaker turn features with the identified significant predictors.
4. Bridge the gap between the model-based generation of non-verbal behaviors (as facial landmarks) and their physical realization by emulating the generated behavior with an embodied agent.
5. Show that our BC smile generation yields appropriate and natural-looking smiles through a user study involving the embodied agent.

Results suggest speaker sex, their use of negations, loudness, word count in the listener's turn, their usage of comparisons, and mean pitch are significant predictors of BC smile intensity. Our generative approach shows that taking listeners' behavior into account improves performance, and adding the conditioning vector offers significant improvements in terms of empirical metrics such as Average Pose Error (APE) and Probability of Correct Keypoints (PCK).

2. Related Work

Existing works have validated the efficacy of an agent-driven conversation in mental health dialogue and counseling situations. DeVault et al. [6], through their agent-based interviews for distress and trauma symptoms, found that participants were comfortable interacting with the agent as well as sharing intimate information. Utami and Bickmore [7] used embodied agents for couples counseling. Participants reported significantly improved affect and intimacy with their partner and generally enjoyed the agent-driven counseling session. Our work builds on this line of research to improve the BC capabilities of agents.

Backchannel behaviors were traditionally produced using a set of predefined rules based on prosodic or linguistic cues of the speaker. Both Ward and Tsukahara [8], Benus et al. [9] have found prosodic cues (particularly pitch and its changes) to be reliable predictors for vocal BC occurrence. In contrast, we use prosody and linguistic cues from both speaker and listener to identify significant predictors of BC smiles.

In the multimodal context, Bertrand et al. [10] studied prosodic, morphological, and discourse markers for their effect on vocal and gestural backchannels (hand gestures, smiles, eyebrows), and Truong et al. [11] explored visual BCs by often limiting them to head nods and, at times, grouping different BCs into the same category [12] without accounting for their intrinsic differences. They depended on the speaker's behavior to identify the occurrence and ignored the listener. In addition to leveraging the listener behavior, we specifically study smiles because of their diversity and include both unimodal (visual) and bimodal (visual together with vocal activity) BC smiles.

Wang et al. [13] introduced diversity in generated smiles by conditioning on a specific class and sampling using a variational autoencoder. Learn2Smile [14] used the facial landmarks of the speaker to generate complete listener behavior by separately predicting the low-

[1]Data and code: https://github.com/bmaneesh/Generating-Context-Sensitive-Backchannel-Smiles/

frequency (nods) and high-frequency (blinks) components of facial motion. Ng et al. [15] leverage the speaker and listener's motion and speech features to predict the listener's future motion information. Unlike earlier works that have been limited to facial expression generation using landmarks, their usage of 3D Morphable Models to define facial expressions offers a flexible solution to generate realistic facial expressions in the presence of diverse head orientations. These solutions focus on the entire listener's behavior and offer no insights about specific BC behaviors. Their integrations are also limited to 3D Morphable Models.

The BC smiles produced in this work not only leverage the speaker and listener activity but also condition the generation on salient factors that were found to be significant predictors of smile attributes – duration (the time elapsed between the onset of a smile and its offset) and intensity (maximum amplitude of a smile). Using an embodied agent, we also bridge the gap between generated landmarks and their physical realization.

3. Dataset

One of the primary challenges in studying non-verbal behavior in mental health interactions is access to an appropriate dataset. Patient-therapist interactions or interactions with mental health professionals are access-restricted to protect the identifiable information of the individuals. As a result, we use a YouTube-based large-scale dataset of face-to-face dyadic interactions–RealTalk [16]. The RealTalk dataset consists of individuals taking turns asking predefined, intimate questions about family, dreams, relationships, illness, and mental health[2]. We believe intimate conversations are among the closest accessible alternatives to studying BC behaviors for mental health applications. In this section, we elaborate on our contributions in terms of the annotations for BC smiles and discuss how they differ by the demographics of the dyads and features from the speaker and listener turn preceding it.

3.1. Annotating Backchannel Smiles

We manually annotated 191 BC smiles from 48 (out of 692) dyadic interactions in the RealTalk dataset. The dyads comprised male and female participants from different ethnicities, and social relationships such as siblings, paternal, romantic, and fraternal. The smiles were nearly balanced across the different interpersonal relationships (see Figure 2). An automated facial expression prediction framework [17] was used to evaluate the reliability of the manual annotations. About 83% (i.e., 158 smiles) of the 191 annotated smiles had an A-level or higher intensity. One outlier smile was dropped because of the extremely long duration. The resultant 157 smiles, along with their predicted intensity, were used in this work. In addition to the video recordings at 25 fps and 720p resolution, the dataset also contains speaker-identified turn-level text obtained through automatic transcription [18]. The individuals in the dyadic interaction occupied fixed positions (left and right) in the videos. In this work, the biological sex of the participants was inferred from the videos. Videos where sex could not be established with confidence were discarded.

[2]The original videos can be accessed from https://www.youtube.com/c/TheSkinDeep

		Speaker sex			
		F	M	F	M
Listener sex	F	32	1	10	10
	M	4	4	6	7
	F	27	7	3	20
	M	7	3	16	0

Figure 2: Distribution of speaker and listener sex across different interpersonal relationships in annotated RealTalk dataset. Relationships are color-coded: siblings (pink), friends (orange), paternal (green), and romantic couple (grey).

3.2. Effect of Sex and Relationship on Smile Attributes

Given various interpersonal relationships in the dataset of individuals of both sexes, we compared the mean duration of backchannel smiles across the factors using ANOVA (Table 1) with type-III sum of squares to account for imbalance between males and females. Two-way interactions between sex, and sex and relationship were also included. The ANOVA analysis suggests that the duration of backchannel smiles differs significantly by listener sex and the interaction effect of the listener sex and relationship. A post hoc Tukey revealed that male listeners, when interacting with their siblings (regardless of speaker sex), express longer BC smiles ($p<0.05$).

Similarly, the intensity of smiles marginally differed by the speaker's sex. The post hoc Tukey revealed that the smiles as a response to a male speaker are less intense than a female speaker ($p<0.1$). ANOVA analysis is presented in the appendix as Table 4.

3.3. Effect of Context Cues

Our contextual cues were extracted from prosody and speech features independently derived from the turns of both the speaker and the listener just before the smile onset. Since the speaker's turn continues while the listener backchannels, speaker activity till the onset of the smiles

Table 1
ANOVA of listener sex, speaker sex, and relationship on duration of smile. '*' indicates p<0.05 and '**' indicates p<0.01).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
$sex_{listener}$	1	12.36	12.36	4.59	0.0339 *
$sex_{speaker}$	1	1.29	1.29	0.48	0.4907
$relationship$	3	4.18	1.39	0.52	0.6709
$sex_{listener}$* $relationship$	3	42.80	14.27	5.29	0.0017 **
$sex_{listener}$* $sex_{speaker}$	1	0.90	0.90	0.33	0.5652
$sex_{speaker}$* $relationship$	3	9.70	3.23	1.20	0.3123
Residuals	144	388.03	2.69		

was considered in this study. The audio was trimmed to the onset to obtain corresponding contextual cues, and the Montreal Forced Aligner (MFA) [19] was used to extract corresponding transcription information.

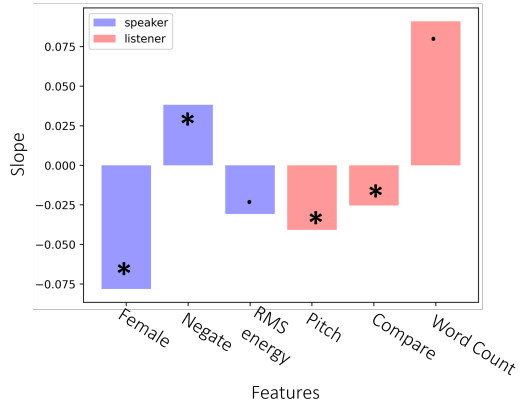

Figure 3: Regression slopes showing the effect of context cues on the intensity of BC smiles. A positive slope indicates the smile intensity increases with a given feature (vice-versa for a negative slope). * indicates slope is significant at p<0.05 and · indicates marginal significance at p<0.1.

Prosody cues: Our prosodic features consisted of some of the fundamental characteristics of speech, such as mean pitch during the turn, range of the pitch, and Root Mean Square (RMS) energy of the audio signal. These features were chosen because of their relevance (see related work) in BC behavior and also due to the ease of interpretation as well as their ability to convey various behavioral traits. For example, RMS energy conveys traits such as confidence, doubtfulness, and enthusiasm [20]. Lastly, using the OpenSMILE [21] software, prosodic features were obtained.

Speech cues: The spoken content of speaker and listener turns was also accounted for through variables from the Linguistic Inquiry and Word Count (LIWC) [22] framework. These variables were word count, usage of negations (no, not, never), comparisons (greater, best, after), interrogative words (how, when, what), valence of the turns (positive or negative emotion), and focus on events in the past, present and future.

A generalized linear model predicted the smile intensity from context cues and dyad demographics. Results using an inverse link function (model explained variance $R^2 = 0.243$) with the prosody and speech cues from the audio signal are presented as Figure 3. Note that the speakers' and listeners' context cues were Z-score normalized. Speaker characteristics such as sex and negations were found to be significant predictors of intensity. Female speakers elicited significantly narrower smiles from their listeners, but the speaker's usage of negations resulted in wider smiles. The speaker's loudness (RMS energy) had a marginally significant negative correlation with the smile intensity. Listener behavior also significantly impacted their BC smiles. Using comparative words by the listener and their mean pitch in their preceding turn resulted in significantly narrower smiles. In contrast, their word count had a marginally significant positive correlation with intensity. A similar analysis for duration did not reveal any significant correlations.

4. Modeling Smiles

To automatically generate BC smile and non-smile activity in listeners, we use the audio from the speaker's current turn and the listener's last turn as input. 15 smiles were dropped due to difficulties in the preprocessing steps with MFA. The remaining 142 annotated smile instances were augmented with an equal number of non-smile instances. The non-smile instances were identified so that they were at least two seconds away from the onset of the closest smile instance, a strategy adopted from [23] for turn-taking prediction. The mean duration of smiling and non-smiling instances was ensured to be the same.

Attention-based generative model: The generative model (Figure 4) for facial landmark prediction primarily consisted of an encoder and a decoder with a one-layer GRU each. Inputs to the model were embeddings from speaker and listener turns extracted using the pretrained vggish model [24]. We limited the input context length to use turn durations of 60 seconds. The output context was limited to predicting one second of facial activity. The speaker vggish embeddings were used as input to the encoder. The hidden state of the GRU was initialized as the mean of the listener's turn embeddings. The fi-

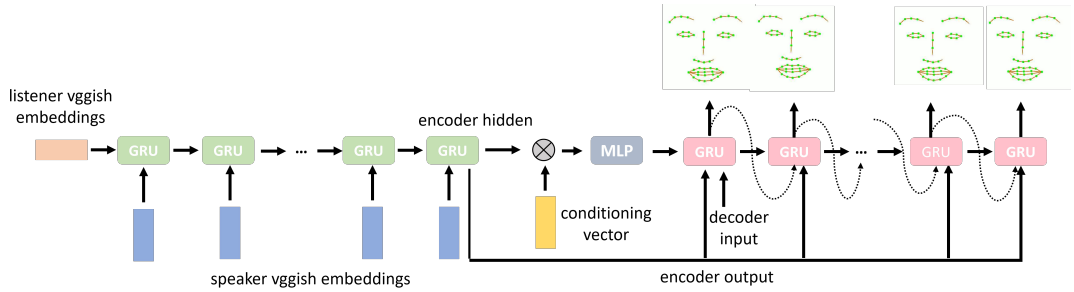

Figure 4: Generative model architecture. Encoder input contains speech embeddings of listener and speaker from the pretrained vggish model. The encoder's final hidden state is concatenated with the conditioning vector and then used to initialize the decoder's hidden state. Decoder output landmarks are sequentially fed (dotted curves) to generate the next landmarks in the output sequence.

nal hidden state of the encoder was concatenated with the conditioning vector, and a linear layer with ReLU activation was used to match the dimensionality of the decoder's hidden state. At each decoding step, attention [25] was applied between the encoder output and the decoder's last hidden state (Equation 1) to use as the input to the next step.

$$a(s_{t-1}, h_i) = v^T tanh(W_a h_i + W_b s_{t-1}) \quad (1)$$

where $a(s_{t-1}, h_i)$ is the attention between decoder last hidden state (s_{t-1}) and encoder output (h_i). W_is and v are linear layers.

4.1. Implementation details

The videos were split into two vertical halves, one corresponding to each individual in the dyadic interaction. These were used for facial landmark extraction using the AFARtoolbox [17]. To account for various facial shapes, we normalized landmarks to the mean face of the dataset using the approach described in [26]. Because of the high degree of correlation between successive frames, frames were downsampled by a factor of three, to use every third frame. Displacement was then calculated as the difference between the landmarks from successive frames. These were further subjected to a min-max normalization to allow for individual differences in smiling dynamics. The normalized displacements were predicted using the attention-based generative model. The predicted frame-level displacements were incorporated into the last known listener facial expression to generate the sequence of facial landmarks recursively.

We enforced teacher-forcing with simulated annealing during training and linearly decreased the likelihood of using ground truth at every 20 epochs. Stochastic Gradient Descent with a learning rate initialized at $1e-4$ weight decay and 0.99 momentum were used to minimize the Mean Squared Error (MSE) between predictions and the ground truth. The learning rate was halved when validation loss plateaued for 20 consecutive epochs. Data was partitioned into 75 (train), 15 (validation), and 15 (test) split in terms of the number of dyads. Models were trained for 250 epochs, and validation loss was used to determine the best model for testing. This was repeated 10 times to evaluate the statistical significance of differences against baseline speaker-based BC generation setting.

Metrics: Objective measures of performance from gesture generation approaches, including Average Pose Error (APE) and Probability of Correct Keypoints (PCK), were adopted to quantify the generated landmarks against the ground truth from the AFAR toolbox. APE (Equation 2) is equivalent to the mean squared error between predicted facial expression and ground truth facial expression. PCK (Equation 3) is a proximity-based metric that considers the landmark to be correctly predicted if the difference with ground truth falls below a margin. We report mean PCK for $\sigma = 0.1$ and 0.2.

$$APE = \frac{1}{k}\sum_{y=1}^{k} \|(\hat{y}(p) - y(p))\|_2 \quad (2)$$

where k is the number of landmarks, $\hat{y}(p)$ is the prediction and $y(p)$ is the groundtruth.

$$PCK_\sigma = \frac{1}{k}\sum_{y=1}^{k} \delta(\|(\hat{y}(p) - y(p))\|_2 \leq \sigma) \quad (3)$$

where δ is an indicator function and σ is the margin.

4.2. Results

Using listener behavior and conditioning vector together with the speaker behavior resulted in improved performance compared to the baseline speaker behavior-based

Table 2
Average Pose Error (APE) and Probability of Correct Keypoints (PCK) metrics for generated facial expressions under various experimental settings. A downward-facing arrow indicates lower value implies better generation. '*' indicates significance with p <0.05 with '·' indicates marginal significance with p <0.1.

Model	APE↓	PCK↑
Speaker only (Baseline)	9.552	0.219
Speaker and Listener	9.346·	0.220·
Speaker and Listener with Conditioning vector	9.279*	0.223*
Speaker and Conditioning vector	9.615	0.218·

prediction. As shown in Table 2, APE decreased by 0.273 points while PCK increased by 0.004; these gains were statistically significant. When listener behavior was added to the speaker behavior, marginally significant improvements were observed. APE reduced by 0.206 points while PCK increased by 0.001 points. These reiterate our hypothesis that both speaker and listener contribute to BC behaviors. When speaker behavior was augmented with the conditioning vector, only nominal differences were observed against the baseline. APE increased by 0.063 points, and PCK decreased by 0.001.

To understand how the performance varies with different smiles, we predicted APE (and PCK) as a linear combination of duration, intensity, and the model configuration using a regression model. Results from Figure 5 show that duration significantly affects the PCK. Interestingly, the positive slope suggests that longer smiles are generated better over shorter smiles. Only a marginally significant effect of duration can be observed for APE. With the increase in the intensity of the smile, the generation performance decreases. This is significant for D-level and E-level smiles. Using listener features and the conditioning vector along with the speaker features improves the performance (negative and positive slopes for APE and PCK, respectively) compared to the baseline speaker-based generation. However, this effect is not statistically significant.

Qualitative evaluation of ground truth landmarks from Figure 6 suggest the deficiencies of the existing facial landmark prediction approaches [17] to accurately track lip corners both in the presence and absence of non-frontal head pose. While a visually noticeable difference can be observed as the smile evolves, the ground truth landmarks fail to capture the subtle lip corner motion. This limitation in the ground truth has resulted in nominal motion in the predicted landmarks. We also found that BC smiles that co-occur with vocal activity are challenging to predict. Figure 7 shows one example where the vertical distance between the upper and lower lips increases and decreases because of the simultaneous *yeah*

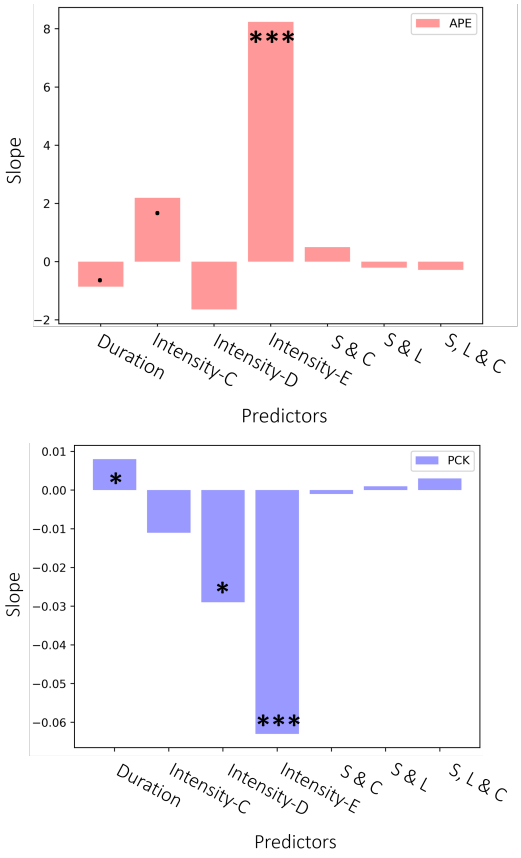

Figure 5: Effect of duration and intensity of smile along with ablation of inputs on generative model performance measured using APE (top) and PCK (bottom). S & C-speaker and conditioning vector, S & L-speaker and listener, and S, L & C-speaker and listener and conditioning vector as inputs to the model. '·', '*' and '***' indicate significance with p <0.1, p <0.05 and p <0.001 respectively.

utterance. However, the model fails to capture this vertical motion.

Metrics like APE and PCK provide an objective measure of the prediction. However, evaluating concepts such as realism and contextual relevance of the BC prediction requires subjective ratings from human evaluation. A convention in evaluating landmark or keypoint-based generative approaches is the human comparison of predicted keypoints against the ground truth [14, 27]. While this might work for problems such as gesture generation that involve a strong motion component, evaluating subtle behaviors like facial expressions using a similar strategy could be challenging. To address this concern, we leverage the emulated version of an embodied agent: Furhat [28].

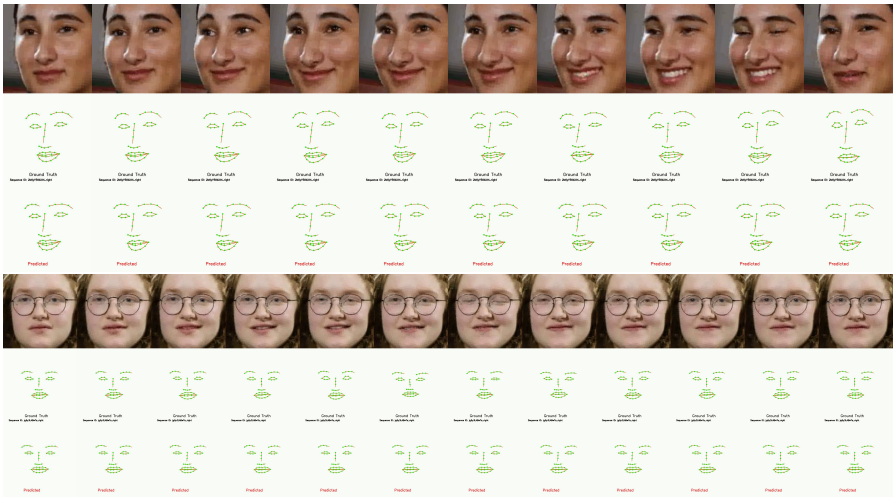

Figure 6: Two sample smiles from the dataset showing their onsets (left-most frame to widest smile frame) and offsets (widest smile frame to right-most frame). Note that while the evolution of smile is noticeable in ground truth landmarks (second row) of the top smile, subtle changes between successive frames of the bottom smile are not captured by its ground truth landmarks. This is also observed in the generated landmarks (third row). Zoom-in recommended. The faces used are from the RealTalk dataset.

Figure 7: Limitation of the current approach in generating a bimodal backchannel smile. The frames highlighted in red box correspond to the co-occurring verbal "yeah". Notice that ground truth landmarks (second row) fail to capture the vertical mouth movement. This is also observed in the generated landmarks (third row). Zoom-in recommended. The faces used are from the RealTalk dataset.

5. Smiles on an Embodied Agent

So far, we have shown modeling smiles by generating facial landmarks. However, users in real-world scenarios do not expect to see such abstract representations of faces. Aligning these facial landmarks with embodied agents is key for an interactable conversational agent. To achieve this, we describe the procedure to transfer generated landmarks to an embodied robotic simulation system called Furhat. We then conduct a user study for subjective perceived differences in Furhat's behavior due to BC smile.

5.1. Emulation Setup

Furhat allows users to control facial expressions using a set of facial parameters called BasicParams[3] (ex. MOUTH_SMILE_LEFT and MOUTH_SMILE_RIGHT to control the left and the right lip corners; BROW_UP_LEFT, BROW_UP_RIGHT to control the left and right eyebrows, etc.). Our setup uses these parameters to enable the embodied agent's smile and express associated eyebrow actions. The landmarks from a generated smile expression were used to calculate the displacement between successive frames and normalized to the [0, 1] range. For eyebrows, only vertical displacement was used. Our inputs to the Furhat API consisted of the lip corner and eyebrow displacements corresponding to the frame with the widest smile (maximum horizontal displacement between the lip corners). The duration of the Furhat smile was set to the duration of the generated smile. Figure 8 shows an example of the resultant expression. The user study was conducted using the Furhat Desktop SDK. However, we do not foresee difficulties transferring the emulation setup to a physically embodied Furhat.

5.2. User Study Procedure

We conducted a small-scale user study of participants watching two pre-recorded videos of the Furhat interacting with an individual. They differ only in terms of Furhat expressing a BC smile. In both interactions, Furhat starts

[3]https://docs.furhat.io/remote-api/#python-remote-api

Figure 8: Four frames of an example Furhat robot emulation with different levels of smiles used as backchannels during the conversation in our user study.

with a brief introduction of itself, followed by a short question–"How have you been feeling over the last two weeks?". As the user responds, a smile is generated at the appropriate location (see Figure 8). We refer to this scenario as the *backchannel* setting. Another video of the same individual interacting with Furhat with no BC (*non-backchannel*) serves as our baseline. Seven graduate students then rated each video recording separately. Note that raters were not primed on the study's outcome, and no explicit instructions about smiles were given.

To quantify the user's perception of Furhat interacting with an individual, the influence of BC smile in addition to the effect of its intensity and duration, and their willingness to interact with one was quantified through the following questions on a 5-point Likert scale (1: strongly agree, 5: strongly disagree).

1. The Furhat's smiles looked human-like.
2. The Furhat's smiles looked natural and friendly.
3. I would talk to this agent frequently.
4. I felt the brightness of Furhat's smiles was appropriate.
5. The Furhat was smiling for longer or shorter duration than it was expected.
6. I would feel comfortable talking to this agent about non-personal topics.
7. I would feel comfortable talking to this agent about personal topics.

In addition, open-ended feedback was also a part of the questionnaire. We believe these questions help identify some user-facing challenges in generating BC behaviors and how they influence users' attitudes to embodied agent-based dialogue systems for conversations related to mental health.

5.3. Results

Table 3 shows that more users (5/7) expressed moderate or higher agreement that the Furhat agent with BC smile was human-like than its counterpart without BC smile (4/7). One user expressed interest in frequently interacting with the agent in backchannel setting while the lack of backchannels resulted in increased hesitancy among users in frequently using it. Three (out of 7) users found that the brightness of the BC smile was appropriate while two found that the duration of BC smile was longer or shorter than expected. While no difference was observed in terms of users' preference for Furhat for personal conversations based on the presence of the BC smile, more users (3/7) responded that they would use Furhat with BC smiles for non-personal conversations over Furhat without BC smiles (2/7).

Table 3
Number of responses that expressed moderate or strong agreement along various factors related to the BC smiles when interacting with Furhat with and without backchannel behaviors.

Question	Backchannel	Non-backchannel
Human-like	5	4
Natural	6	6
Willing to interact	1	0
Appropriate brightness	3	5
Longer or shorter smiles	2	0
Personal conversations	1	1
Non-personal conversations	3	2

6. Discussion

Our quantitative results suggest that both speaker and listener behavior are important in generating BC behavior. Using listener behavior together with the conditioning vector offered statistically significant improvements in performance when compared to the baseline speaker-only model. This effect was observed both in terms of APE and PCK. We also found that our attention-based generative model can predict low-intensity smiles better than high-intensity smiles. Our user study shows that more people find our agent human-like when it was able to express BC smiles. Participants prefer to interact with it over the agent with no BC smile capabilities for non-personal conversations. However, for intimate personal conversations, the presence of a BC smile did not sway their decision.

Some limitations of this work include the following. We employed an affordable measure of reliability for BC smile annotations using a prediction model over a human rater. A robust approach would involve at least one more human annotator to perform reliability annotations on a portion of the dataset. The statistical analysis also assumes that the smiles were independent of the individuals and dyads. However, a given individual typically produces multiple smiles. Grouping of smiles by factors such as individuals and dyads can be better modelled using a mixed-effects model. Our user study was designed to demonstrate the feasibility of transferring generated facial landmarks to an embodied agent together with understanding *perceived* differences between interactions with and without BC smiles. An appropriate evaluation

framework would include the user interacting with the agent. Followed by a comparison of qualitative subjective ratings of user experience and quantified parameters (such as difference in turn duration, language usage, etc.) of the interaction with and without BC smiles. We believe such approaches provide a holistic evaluation to identify critical instances in the interaction. Lastly, we focused on BC smiles leaving out other conventional signals such as vocal and headpose-based BCs, and how they are affected by the cues from the speaker and listener.

7. Conclusion

To enable BCs in embodied agents for mental health applications, we proposed an annotated dataset of face-to-face conversations including topics related to mental health. Our statistical analysis showed that speaker gender together with prosodic and linguistic cues from both speaker and listener turns are significant predictors of the BC smile intensity. Using the significant predictors together with the speaker and listener behaviors to generate BC smiles offers significant improvements in terms of empirical metrics over the baseline speaker-centric generation.

We bridge the gap between conventional non-verbal behavior generation approaches such as landmarks and poses and their realization by showing that generated landmarks can be transferred to an embodied agent. Thus creating the opportunity for evaluation with a human-like manifestation over a traditional evaluation by comparing generated landmark (or keypoint) outputs. Our small-scale user study suggests our Furhat agent that backchannels is more human-like and are more likely to attract users for non-personal interactions. In addition to these contributions, we also discussed some limitations in existing technology towards generating accurate ground truth landmarks through examples such as failure to capture mouth movement in bimodal BCs and how they affect the generated outputs. We believe these limitations also serve as directions for future research. Our work serves as a baseline for computer scientists interested in behavior generation, and an attractive source of BC smiles for behavioral scientists to study the effect of context cues on BC smiles in intimate conversations.

8. Ethical Statement

We proposed a generative approach for backchannel smile production to enable naturalistic interactions with embodied AI agents for mental health dialogue. While our dataset offers diverse smiles from people in different interpersonal relationships, like many existing generative approaches, the choice of pretrained embeddings, imbalance between males and females, lack of male-male romantic relationships, and lack of age and ethnicity information in the dataset might have resulted in biased generations. We also acknowledge that using embodied agents in such sensitive applications should undergo rigorous evaluations by technical and domain experts and regulatory bodies. In our work, we do not interpret embodied agents as a substitute for professionals in mental health or allied areas of healthcare but to provide tools for them to better serve the community's demands. We believe that the advantages and limitations of embodied agents in mental health should be presented to the users and the healthcare experts to provide maximum benefits. The information used in this work is identified from a publicly available dataset. Also, special attention has been paid to privacy and copyright requirements for relevant images showing individual faces. The user study raters were voluntary participants, and the University of Pittsburgh IRB approved the data collection.

9. Acknowledgments

Bilalpur and Cohn were supported by the U.S. National Institutes of Health through award MH R01-096951. Zeinali was supported through the Khoury Distinguished Fellowship at Northeastern University.

References

[1] H. Modi, K. Orgera, A. Grover, Exploring barriers to mental health care in the u.s. (2022). doi:10.15766/rai_a3ewcf9p.

[2] S. Song, S. Jaiswal, L. Shen, M. Valstar, Spectral representation of behaviour primitives for depression analysis, IEEE Transactions on Affective Computing 13 (2020) 829–844.

[3] F. Ceccarelli, M. Mahmoud, Multimodal temporal machine learning for bipolar disorder and depression recognition, Pattern Analysis and Applications 25 (2022) 493–504.

[4] Y. Yang, C. Fairbairn, J. F. Cohn, Detecting depression severity from vocal prosody, IEEE transactions on affective computing 4 (2012) 142–150.

[5] Z. Ambadar, J. F. Cohn, L. I. Reed, All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous, Journal of nonverbal behavior 33 (2009) 17–34.

[6] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al., Simsensei kiosk: A virtual human interviewer for healthcare decision support, in: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, 2014, pp. 1061–1068.

[7] D. Utami, T. Bickmore, Collaborative user responses in multiparty interaction with a couples counselor robot, in: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2019, pp. 294–303.

[8] N. Ward, W. Tsukahara, Prosodic features which cue back-channel responses in english and japanese, Journal of pragmatics 32 (2000) 1177–1207.

[9] S. Benus, A. Gravano, J. B. Hirschberg, The prosody of backchannels in american english (2007).

[10] R. Bertrand, G. Ferré, P. Blache, R. Espesser, S. Rauzy, Backchannels revisited from a multimodal perspective, in: Auditory-visual Speech Processing, 2007, pp. 1–5.

[11] K. P. Truong, R. Poppe, I. de Kok, D. Heylen, A multimodal analysis of vocal and visual backchannels in spontaneous dialogs., in: INTERSPEECH, 2011, pp. 2973–2976.

[12] A. Gravano, J. Hirschberg, Backchannel-inviting cues in task-oriented dialogue, in: Tenth Annual Conference of the International Speech Communication Association, 2009.

[13] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, N. Sebe, Every smile is unique: Landmark-guided diverse smile generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7083–7092.

[14] W. Feng, A. Kannan, G. Gkioxari, C. L. Zitnick, Learn2smile: Learning non-verbal interaction through observation, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 4131–4138.

[15] E. Ng, H. Joo, L. Hu, H. Li, T. Darrell, A. Kanazawa, S. Ginosar, Learning to listen: Modeling non-deterministic dyadic facial motion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20395–20405.

[16] S. Geng, R. Teotia, P. Tendulkar, S. Menon, C. Vondrick, Affective faces for goal-driven dyadic communication, arXiv preprint arXiv:2301.10939 (2023).

[17] I. O. Ertugrul, L. A. Jeni, W. Ding, J. F. Cohn, Afar: A deep learning based tool for automated facial affect recognition, in: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), IEEE, 2019, pp. 1–1.

[18] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised pre-training for speech recognition, arXiv preprint arXiv:1904.05862 (2019).

[19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi, in: Proc. Interspeech 2017, 2017, pp. 498–502. doi:10.21437/Interspeech.2017-1386.

[20] S. A. Memon, Acoustic correlates of the voice qualifiers: A survey, arXiv preprint arXiv:2010.15869 (2020).

[21] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

[22] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of LIWC2015, Technical Report, 2015.

[23] E. Ekstedt, G. Skantze, Voice activity projection: Self-supervised learning of turn-taking events, arXiv preprint arXiv:2205.09812 (2022).

[24] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., Cnn architectures for large-scale audio classification, in: 2017 ieee international conference on acoustics, speech and signal processing (icassp), IEEE, 2017, pp. 131–135.

[25] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[26] S. Stoll, N. C. Camgöz, S. Hadfield, R. Bowden, Sign language production using neural machine translation and generative adversarial networks, in: Proceedings of the 29th British Machine Vision Conference (BMVC 2018), British Machine Vision Association, 2018.

[27] C. Ahuja, D. W. Lee, R. Ishii, L.-P. Morency, No gestures left behind: Learning relationships between spoken language and freeform gestures, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1884–1895.

[28] S. Al Moubayed, J. Beskow, G. Skantze, B. Granström, Furhat: a back-projected human-like robot head for multiparty human-machine interaction, in: Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers, Springer, 2012, pp. 114–130.

10. Appendix

10.1. Distribution of Intensity and Duration of Smiles

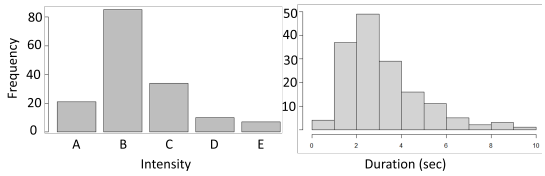

Figure 9: Distribution of intensity and duration of BC smiles in the annotated dataset. The spread of the histograms shows the diversity of the annotated smiles.

Figure 9 shows the distribution of annotated Backchannel (BC) smiles in terms of their intensity and duration. The predicted intensity using the automated approach showed that over 50% of smiles were of B-level intensity, and fewer instances of high-intensity smiles (D and E-levels) were also present. The mean duration was 3.18 ± 1.71 seconds.

10.2. Effect of Sex and Relationship on Smile Intensity

Table 4
ANOVA of listener sex, speaker sex, and relationship on intensity of smile. '·' indicates significant at p<0.1.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
$sex_{listener}$	1	0.53	0.53	0.60	0.4417
$sex_{speaker}$	1	2.93	2.93	3.31	0.0710 ·
$relationship$	3	3.23	1.08	1.22	0.3055
$sex_{listener}$* $relationship$	3	2.00	0.67	0.75	0.5225
$sex_{listener}$* $sex_{speaker}$	1	0.10	0.10	0.11	0.7424
$sex_{speaker}$* $relationship$	3	3.15	1.05	1.19	0.3176
Residuals	144	127.49	0.89		

Note that the intensity of the smile differs marginally by the speaker sex. It is not affected by other factors such as relationship, listener sex and their interaction.