LUCIA FANG

BACKCHANNEL HCI LIT REVIEW

LUCIA FANG

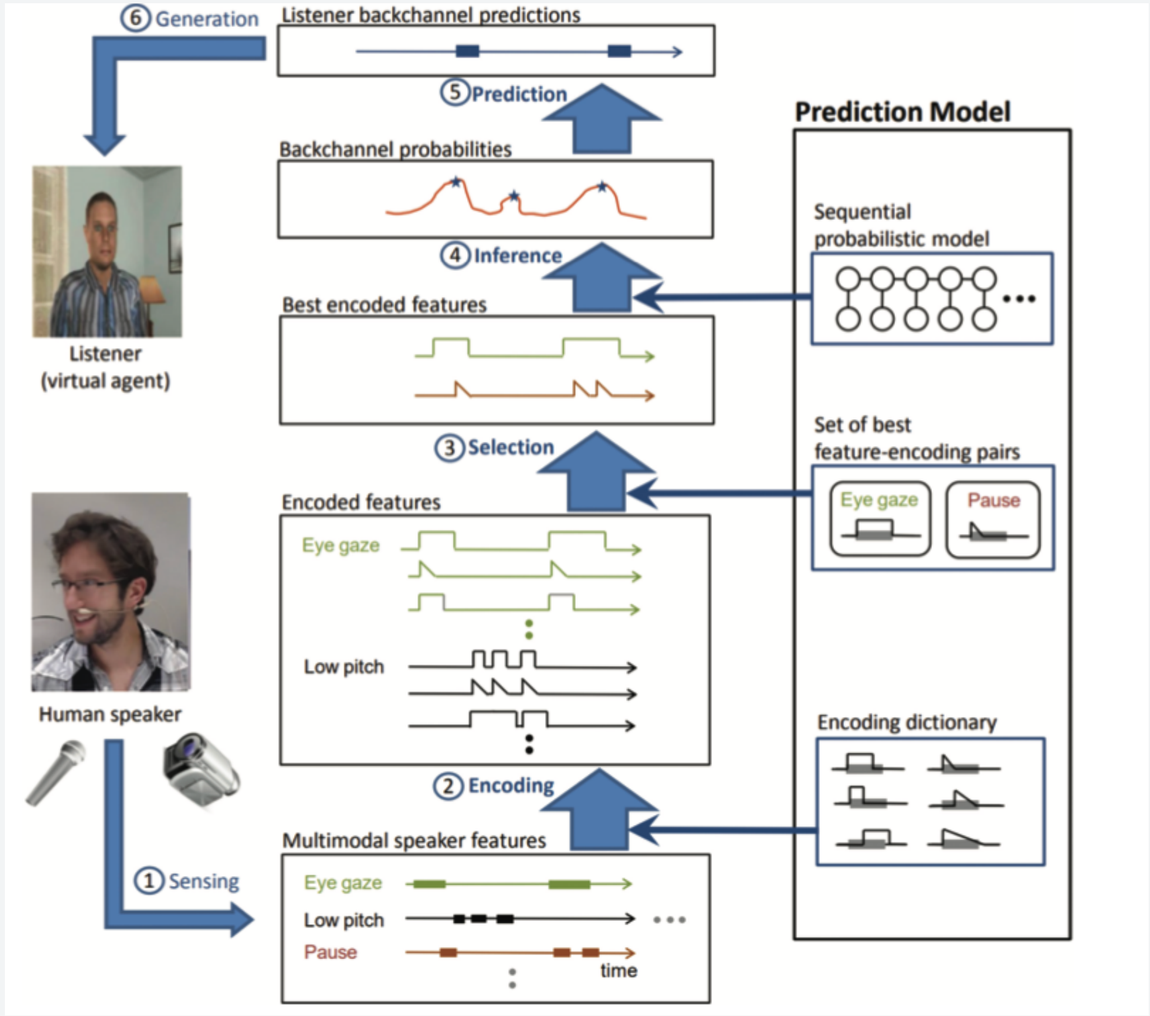
Backchannels

THE VERBAL AND NONVERBAL CUES THAT LISTENERS PROVIDE TO INDICATE THEIR ENGAGEMENT

(Gratch et al., 2021)

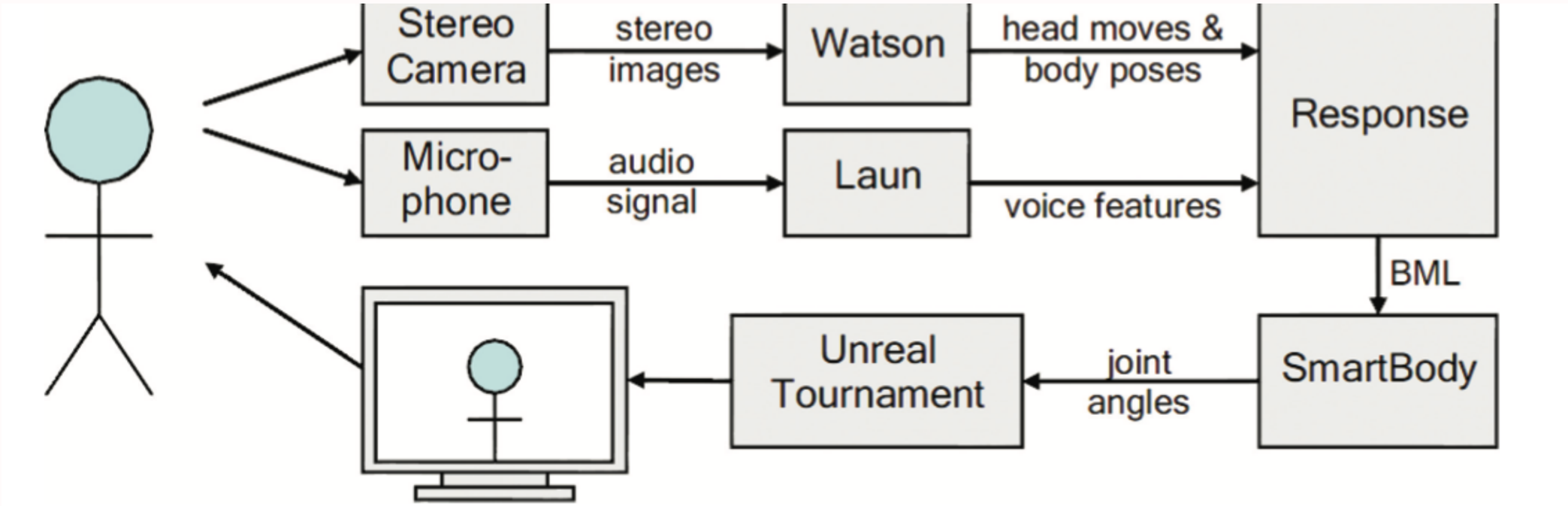
Rapport between humans and socially interactive agents

- Record speech + images from the speaker, use machine learning algorithms to isolate details of face movement or speech/no speech.
- What are the most important features (performance guided, to reduce non-relevant features) and use those to build a classifier --> (BC/noBC).
- Once you have a classifier built, the action is to have the socially interactive agent speak.



(Gratch et al., 2021)

Rapport Agent architecture



(Gratch et al., 2021)

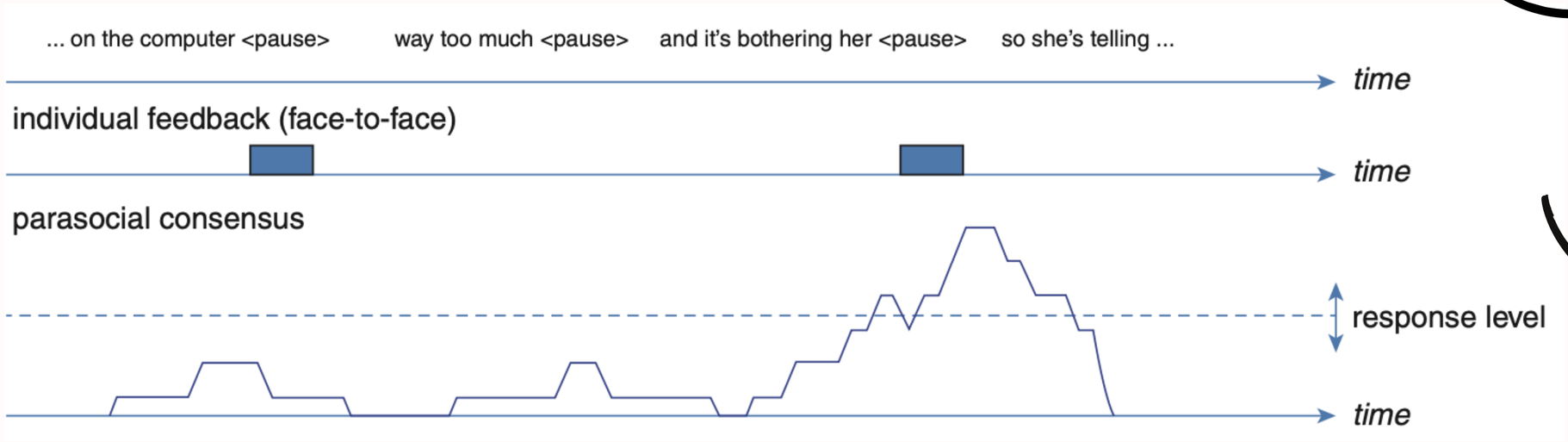
Parasocial interaction:

watch a storyteller and “act out” the types of responses

Parasocial consensus sampling:

Different coders reach a consensus response

(Gratch et al., 2021)



The first line shows the speaker's transcript.

The second line shows the backchannels provided by the original "real" listener.

The bottom line shows the distribution of responses by parasocial consensus sampling coders.

(Gratch et al., 2021)

SimSensei: Automatic assessment of depression, anxiety, PTSD

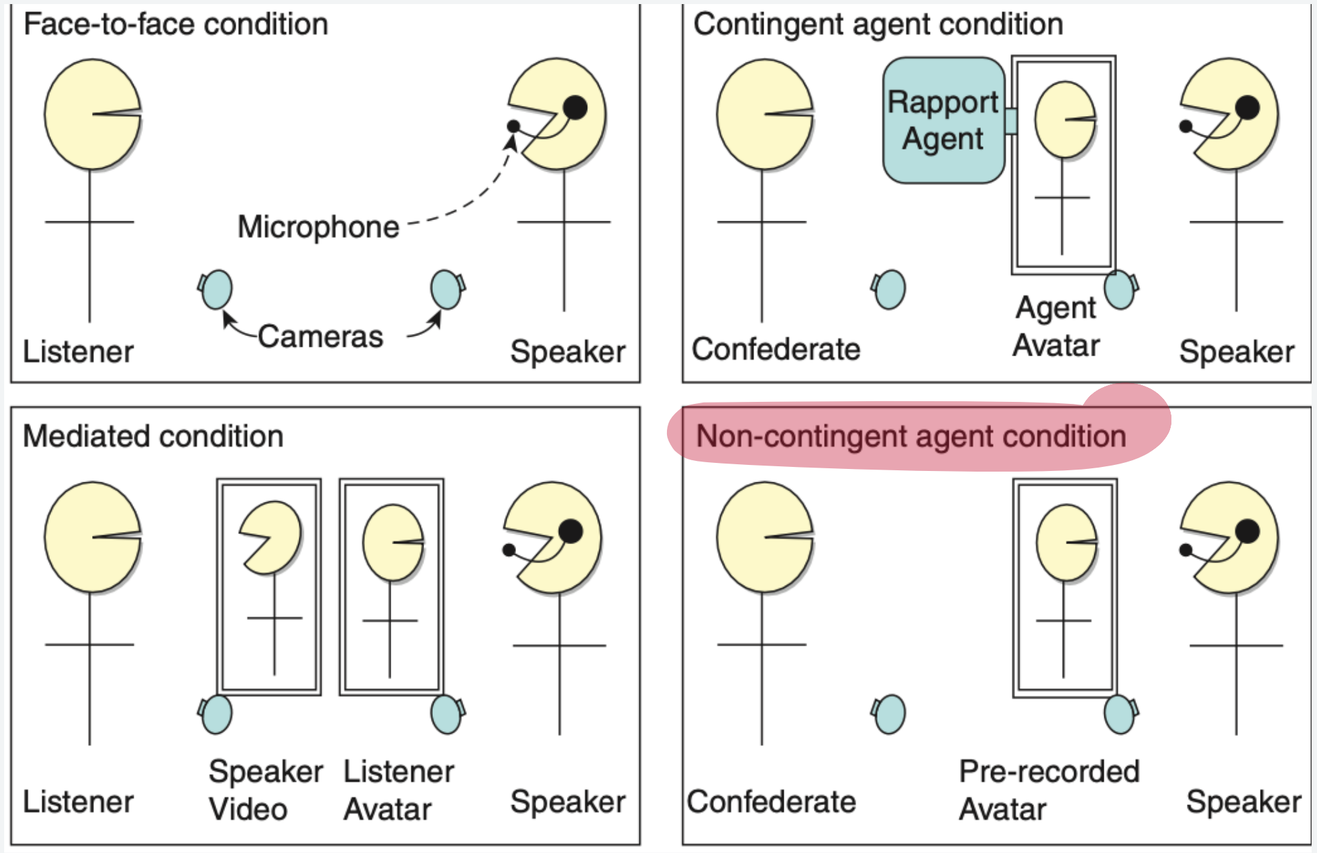
1. Initial chitchat designed to enhance rapport
2. Diagnostic questions
3. Active dialog management allowed the agent to ask follow-up ques tions.
4. Verbal and valanced backchannel feedback
    a. “that’s great” or “I’m sorry”
    b. nonverbal feedback
5. Reduce participants’ fear of being judged and increased comfort

(Gratch et al., 2021)

Experimental paradigm overview:

Bottom right:

- **Make non-contingent agent condition similar contingent agent condition**
- **Body movement/realistic avatar/virtual reality(3D)**



(Gratch et al., 2021)

Factors that influence rapport/interactiveness:

- Positivity: head nods
- Coordination: behaviors are timely and positive, correct facial expression
- Mutual attentiveness: Eye gaze should not stare continuously without nodding
- Anonymity: good for mental health

(Gratch et al., 2021)

Introverts need a lot positive and coordinated nonverbal feedback.

(Gratch et al., 2021)

Find the common backchannels described across all human observers, and use only the common ones for labels.

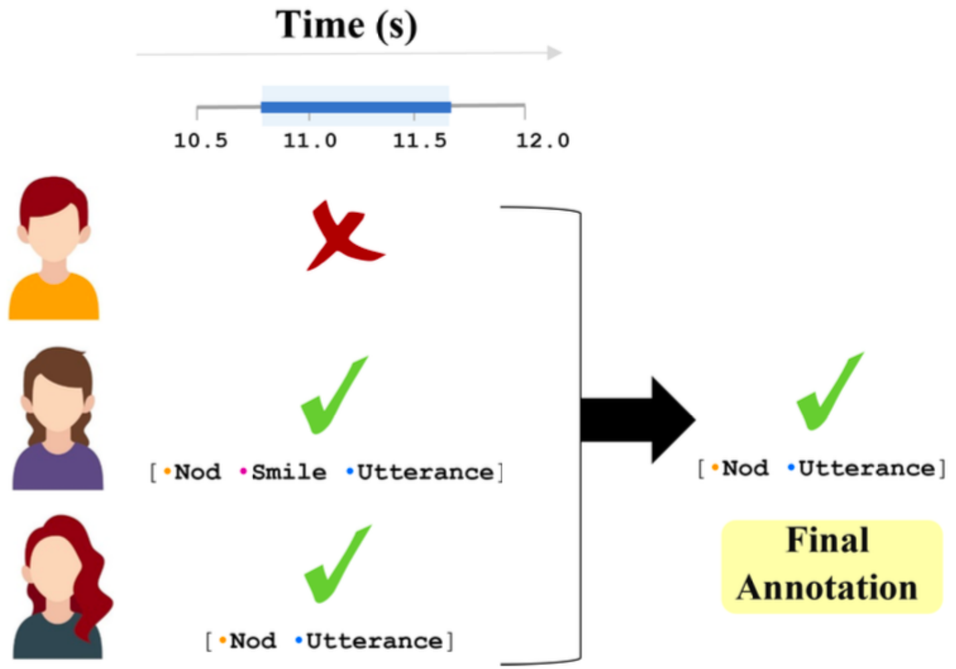


Figure 1: Sample depicting the consensus strategy adopted for combining the annotations from different coders.

(IIIT) et al., 2021)

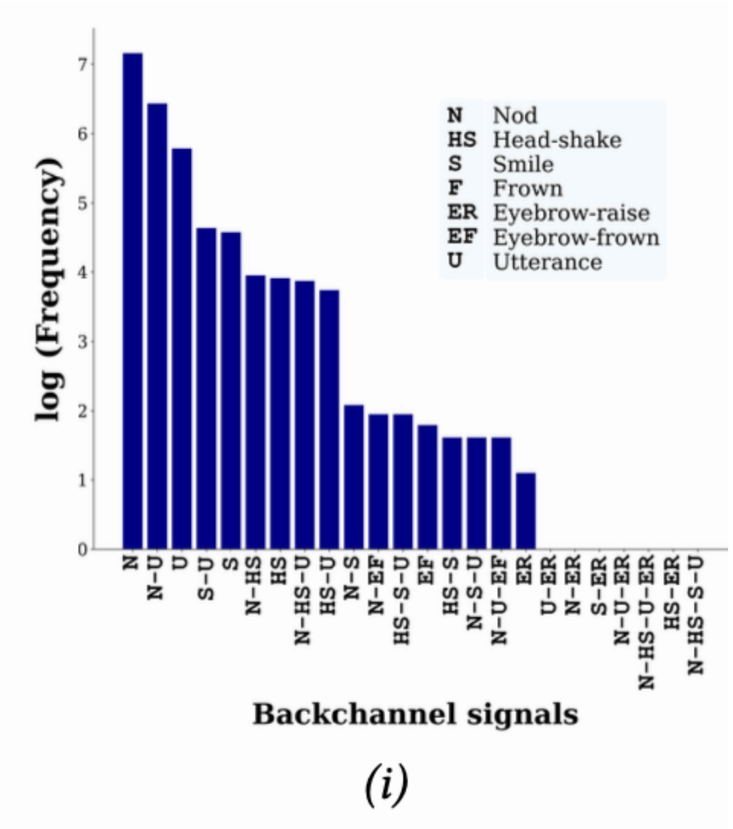
EDA

BC Signal	Labels (*Default*)	N	Mean Freq.	Mean Dur. (s)
Nod	*none*, nod	2037	42.4	1.7
Head-shake	*none*, head-shake	207	4.3	1.7
Mouth	*neutral*, smile/laugh, frown	227	4.7	1.8
Eyebrow	*neutral*, raise, frown	27	0.6	2.1
Utterance	*none*, short-utterance (eg: *"ohh", "okay"*)	1161	24.2	1.4

the average number of times a participant emits a particular signal during a conversation

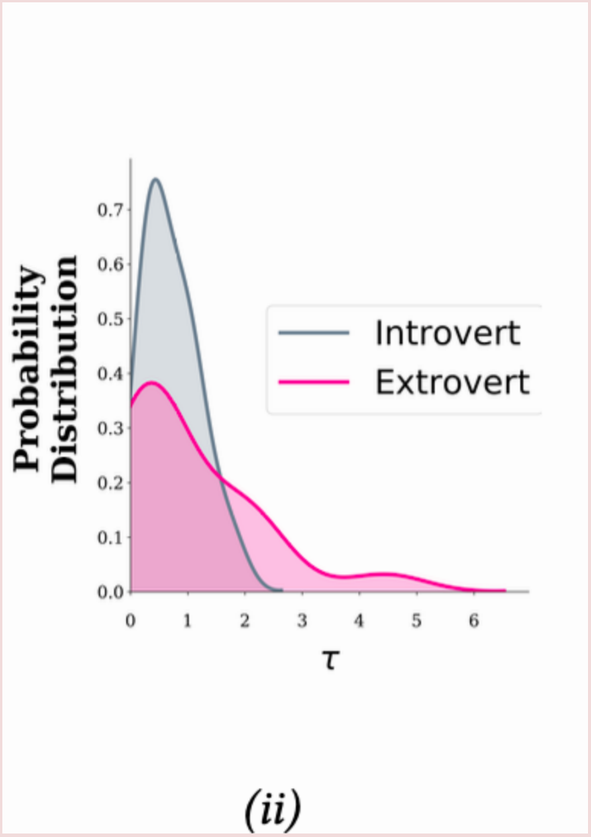
the average duration of the signals in seconds (s)

(IIIT) et al., 2021)



(i)

frequent multimodal instances included at least either a nod or an utterance



(ii)

τ as the ratio of the number of multimodal to unimodal backchannel instances emitted by a subject

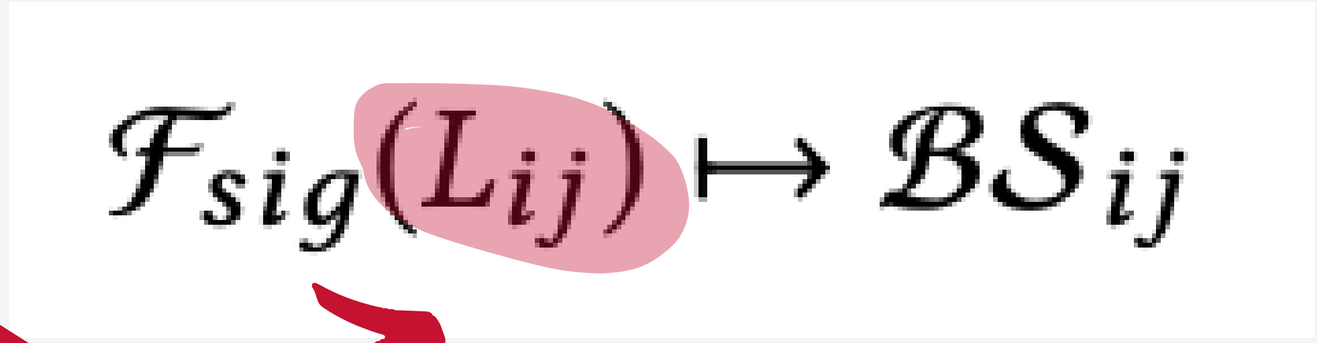
Introverts: 0.35% use multimodal, 0.65% use unimodal

Extroverts: 0.51% use multimodal, 0.49% use unimodal

(IIIT) et al., 2021)

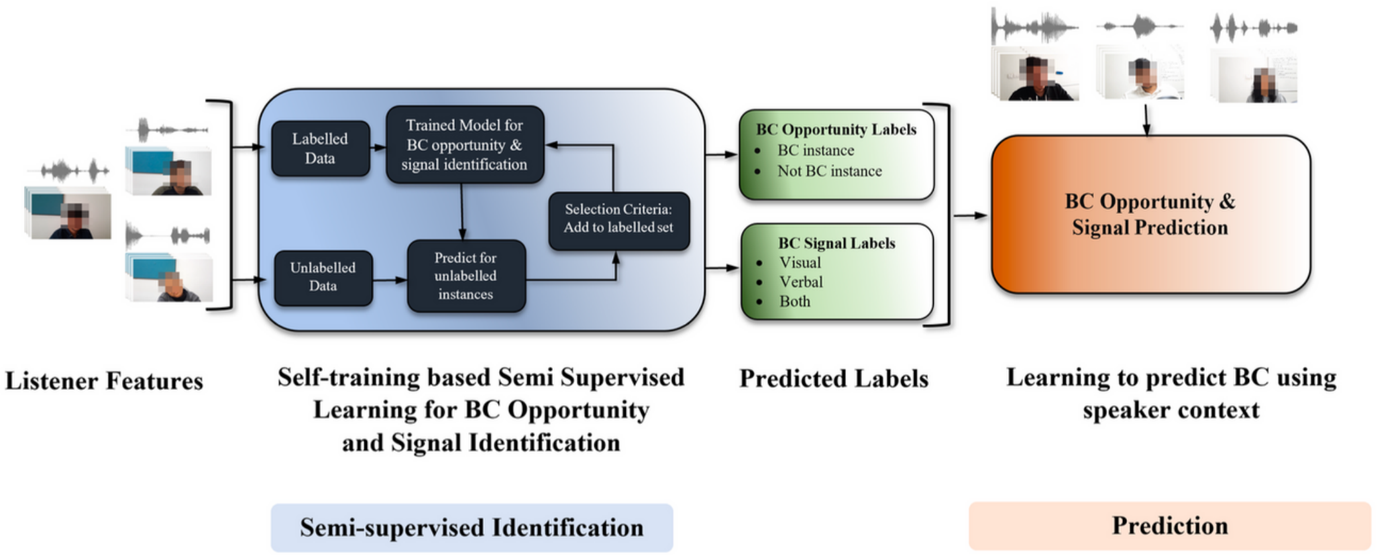
Semi-supervised BC Task Formulation

listener's visual and acoustic features from ith second to jth second



Mapping a time series in the listener's feature space to the corresponding backchannel opportunity label (Verbal/ Visual or Both)

(IIIT) et al., 2021)



backchannels and type of signals emitted using a subset of labeled data

predict these instances and signals using the speaker's context

1. **Extract the listener features and map those to listener backchannels (nod, utterance, smile)**
2. **After mapping, extract features from the speaker side, and map those to predicted listener backchannels.**
3. **Then, deploy this model within a socially interactive agent, which could perform backchannels given speaker context.**

(IIIT) et al., 2021)

These feature extraction codes are all open source python modules. And the classifier uses open source tensorflow LSTM.

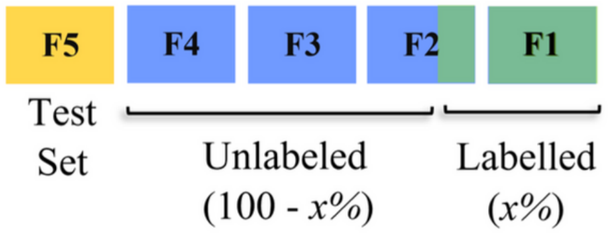


Figure 4: 5-Fold evaluation of semi-supervised models for backchannel opportunity and signal identification.

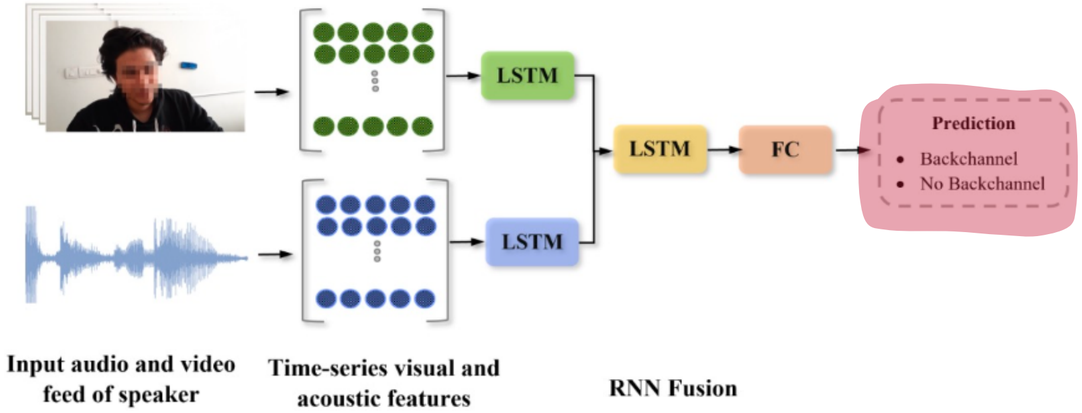


Figure 5: Multimodal RNN Fusion based architecture for backchannel opportunity prediction model.

(IIIT) et al., 2021)

1

Feature set	Precision	Recall	F1-score	Accuracy
Video	0.61	0.86	0.71	0.66
Audio	0.64	0.90	0.74	0.70
Video + Audio	**0.66**	**0.89**	**0.75**	**0.72**

Performs the best!

2

Labels used	Precision	Recall	F1-score	Accuracy
Supervised	0.66	0.89	0.75	0.72
Semi-Supervised	0.62	0.82	0.70	0.66

Table 5: Backchannel opportunity prediction: comparison of model trained using supervised manually-annotated labels with the one using labels generated by the identification module.

(IIIT) et al., 2021)

		Predicted		
		C1	C2	C3
True	C1	0.71	0.07	0.23
	C2	0.07	0.87	0.06
	C3	0.27	0.05	0.67

(*i*)

		Predicted		
		C1	C2	C3
True	C1	0.70	0.09	0.21
	C2	0.08	0.80	0.12
	C3	0.23	0.06	0.72

(*ii*)

Table 7: Confusion matrices for the best backchannel signal prediction model (*i*) trained on manually-annotated labels, (*ii*) trained using labels generated by the signal identification models. Here, C1, C2, C3 refer to the 'visual', 'verbal', and 'both' class labels, respectively.

C2 (Verbal) outperformed C1 (Visual), and both outperformed C3 (both)

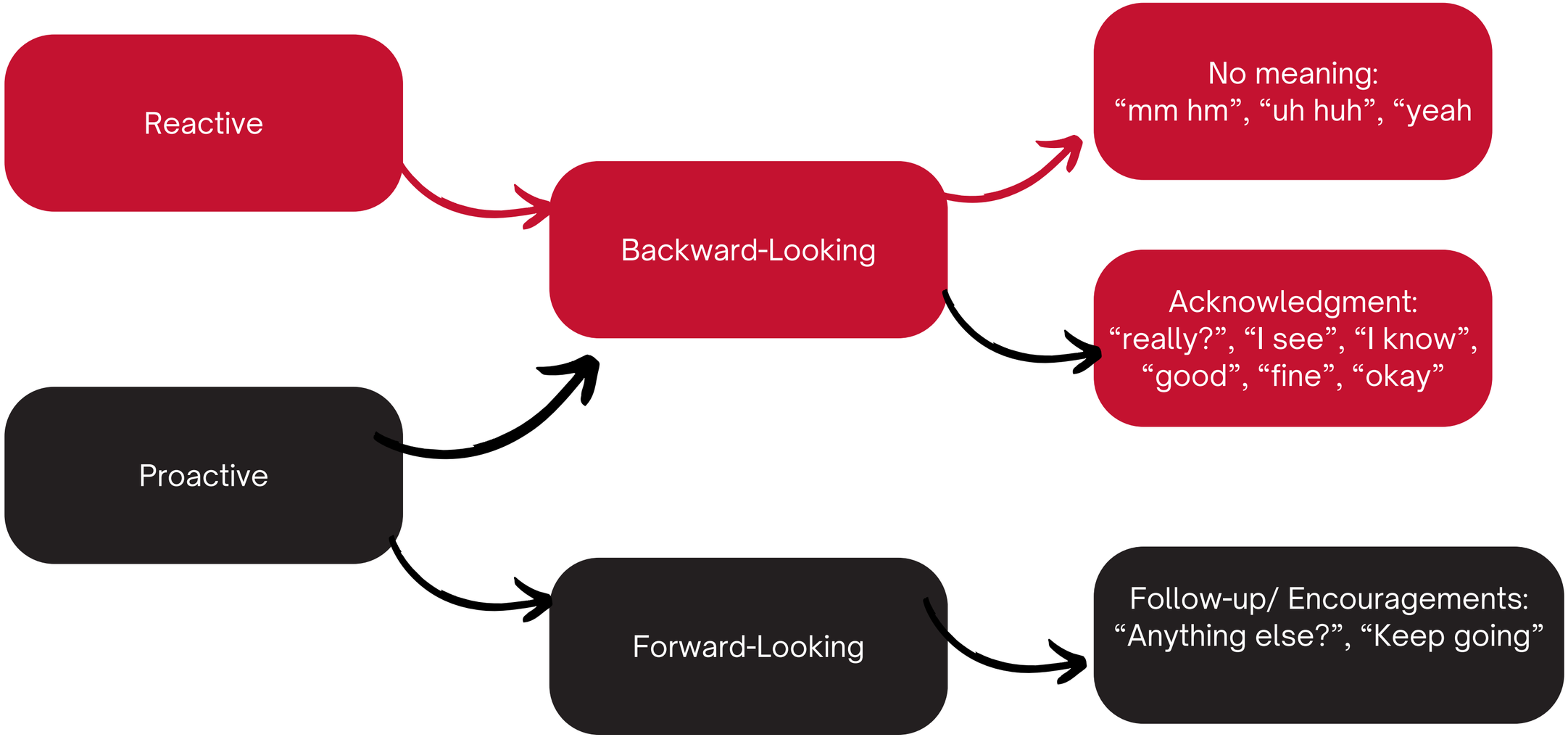
- Within this backchannel classifier, it seems to mistake visual with both visual and audio --> socially interacting agents might do both utterance and head nod/smile when it's supposed to just head nod/smile, and vice versa.

- When the socially interacting agents only needed to utter, it seems to perform quite well (~90%).

- Visual dataset is too small (Eyebrow N=27)

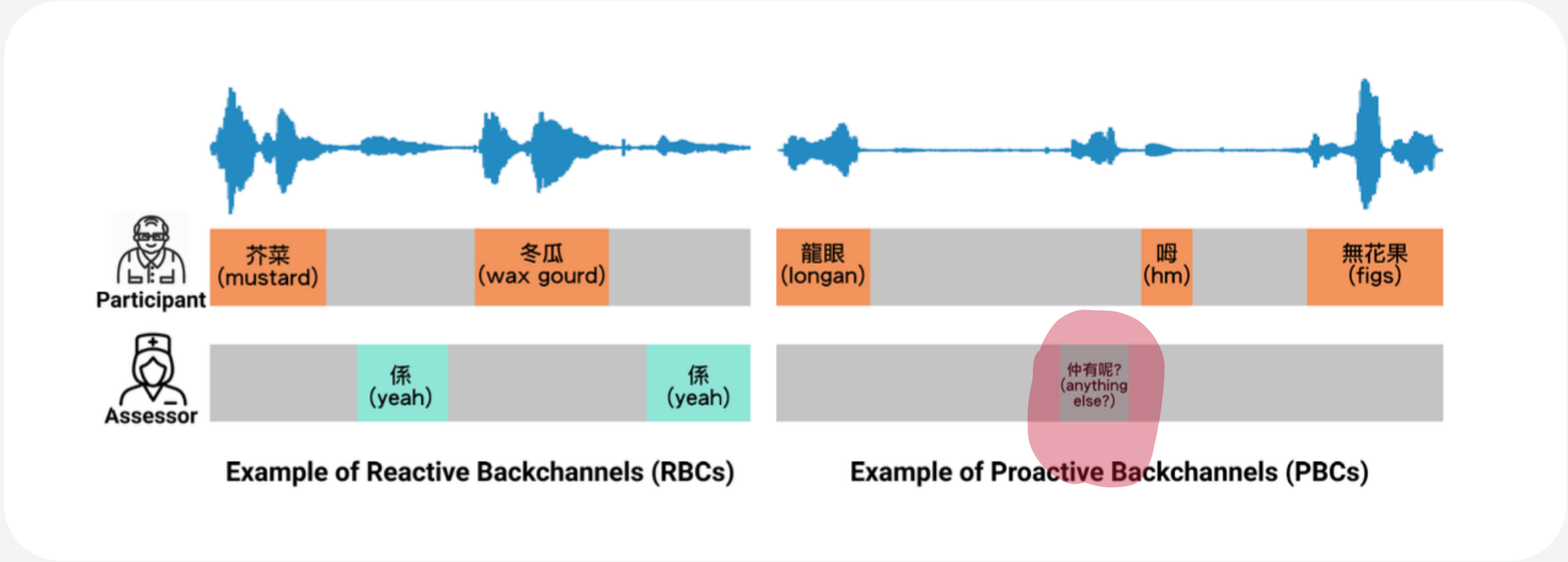
(IIIT) et al., 2021)

Reactive vs Proactive



(Zijian Ding College of Information Studies et al., 2022)

Visual timeline representation of pro vs reactive:



(Zijian Ding College of Information Studies et al., 2022)

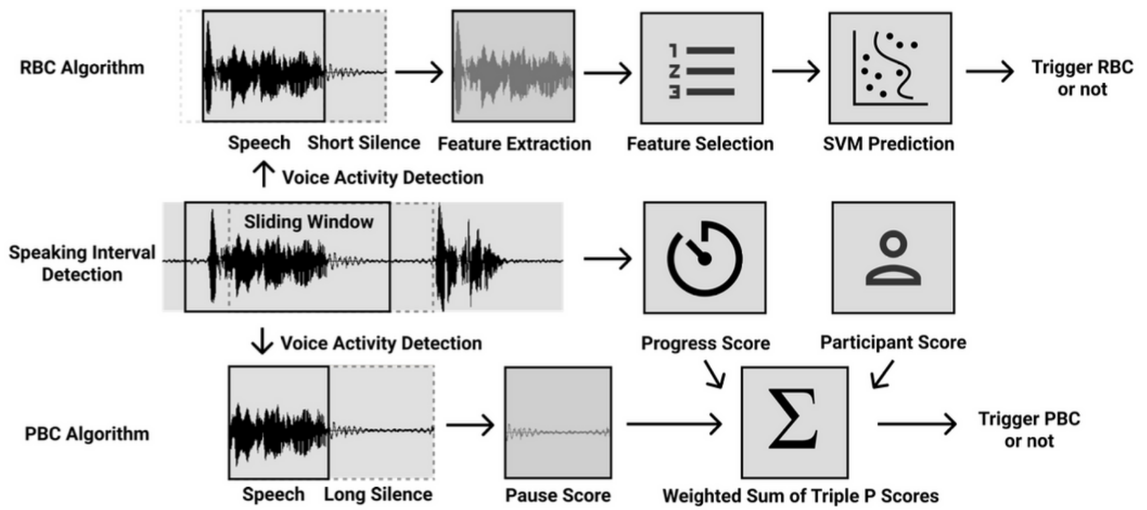


Figure 5: Overview of the pipeline for processing speech data and generating RBC and PBC decisions.

RBC & PBC: feature extraction, feature selection, and a machine learning classifier to predict if backchannel or not

PBC

- Pause Score
  - Range of (0, 1)
  - Increases as the participant's pause becomes longer
- Progress Score
  - Range of (0, 1)
  - Measured task-level timing, indicating when PBCs should occur within a task.
- Participant Score
  - Adjusted the proactive level to adapt to different participants
- Overall PBC Score

$$Score_{PBC} = w_{pau} * Score_{pau} + w_{pg} * Score_{pg} + w_{pt} * Score_{pt}$$

(Zijian Ding College of Information Studies et al., 2022)

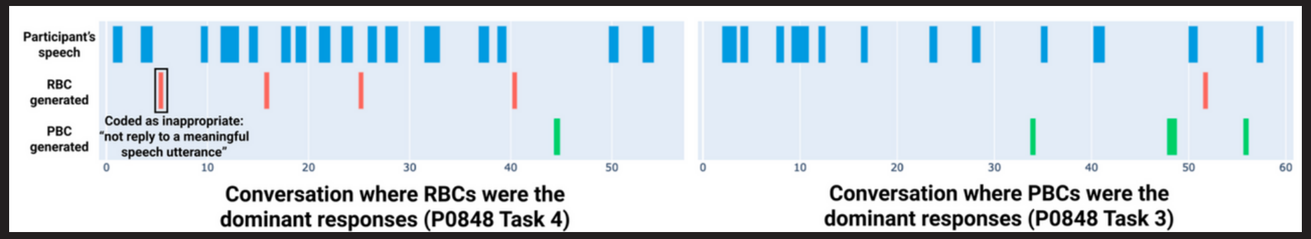
TalkTive:
a React frontend + Flask python server

Introduce the task by playing an audio clip of task description(can be replayed)

Within a time limit, user responds to the prompt by clicking "Start answering”.

If a RBC or a PBC was triggered, TalkTive system would randomly play a piece of corresponding backchannel audio as feedback.

(Zijian Ding College of Information Studies et al., 2022)



PBCs had a slightly higher chance than RBCs to be regarded as inappropriate, which aligned with the proactive nature of PBCs that they were encouraging but had the risk of being aggressive.

1. “Not reply with a meaningful speech utterance” --> RBCs "uhmm"
2. “Urged the speaker” --> PBCs were triggered within a short period of time.
3. “Interrupted speaker’s thinking” --> Both in Open-ended self-disclosure

(Zijian Ding College of Information Studies et al., 2022)

Proactive backchannels could be very important, especially for older populations [Survey]

1. Older adults reported that receiving only RBCs was not as good as gaining responses from humans (Hearng impairments)

2. PBCs were appreciated by older adults, especially when they ran out of answers and were about to give up.

"There was no feedback. I noticed responses like 'hmm', but I knew it's just recording. I didn't think it's real. Human assessors would be better. For human assessors I could see their facial expressions to know whether they were paying attention." (P0215)

said, "While I was thinking, (it said) 'uhm', 'keep going', so I would really think (about the answers), much better than if it didn't give any response...because it made me feel like that there was really a person there, 'take your time', not like facing a cold computer".

(Zijian Ding College of Information Studies et al., 2022)

Future Design Considerations

1

- Situate participants in an appropriate arousal level and engage them in the task without making them feel too stressful
- PBCs can be more effective.

2

- Resolve the tension between encouraging versus pushing the participants while they perform intellectually demanding tasks

3

- Improve participant-related adaptivity of backchan- neling strategies

(Zijian Ding College of Information Studies et al., 2022)