

---

# Designing Emotional Expressions of Conversational States for Voice Assistants: Modality and Engagement

**Yang Shi**

Tongji University  
Shanghai, China  
yangshi.idvx@tongji.edu.cn

**Yongqi Lou**

Tongji University  
Shanghai, China  
lou.yongqi@gmail.com

**Xin Yan**

Tongji University  
Shanghai, China  
xyan330@tongji.edu.cn

**Nan Cao**

Tongji University  
Shanghai, China  
nan.cao@gmail.com

**Xiaojuan Ma**

Hong Kong University of Science  
and Technology  
Hongkong, China  
mxj@cse.ust.hk

**Abstract**

The use of voice-activated virtual assistants (VAs) to execute tasks, request information, or entertain oneself on the smartphone is on the rise. However, insufficient feedback on the states of VAs may impair the interaction flow. We propose to use nonverbal emotional expressions to indicate a VA's conversational states and promote user engagement. We introduce three emotional expression designs of VA, iconic facial expressions, a text box with body movements, and emotional voice waveforms. Our user study results show that a VA with an expressive face or text body movements can evoke stronger user engagement than the one with voice waveforms.

**Author Keywords**

Emotional Expression Design; Voice Assistants; Conversational States; User Engagement.

**ACM Classification Keywords**

H.5.2 [Information interfaces and presentation (e.g., HCI)]: User Interface

**Introduction**

The use of voice-activated virtual assistants (VAs) on smart devices, such as Siri and Google Assistant, is on the rise. Many users treat VAs as social actors during the conversations [12], since they can exhibit various forms of human-

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.  
*CHI'18 Extended Abstracts, April 21–26, 2018, Montréal, QC, Canada.*  
ACM ISBN 978-1-4503-5621-3/18/04.  
<http://dx.doi.org/10.1145/3170427.3188560>

**Conversational States:**

**S1: Standby.** The VA is ready to receive and process inputs from users.

**S2: Listening.** The VA is listening until the user completes the utterance.

**S3: Receiving.** The VA has received the user's words, but may miss some parts to parse the entire utterance.

**S4: Parsing.** The VA has identified the user's utterance, but may require further information to interpret it.

**S5: Interpreting.** The VA has reached an interpretation, but may not be able to map it onto an application.

**S6: Intending.** The VA has mapped the utterance onto a command. It may require the user's permission to act.

**S7: Acting.** The VA is running the command.

**S8: Reporting.** The VA is reporting about the success or failure in executing the command.

like behaviors in their speech, such as wit and sarcasm. However, to maintain a good interaction flow, it is also important for VAs to show nonverbal signals of the current conversational state [1]. For one thing, such signals may foster users' *cognitive engagement* during the interaction, that is, promoting their concentration on the ongoing task [13]. For another, being expressive may have an additional benefit on VAs. It can create a sense of embodiment, which allows users to apply readily available communication strategies established in Human-Human Interaction. As a result, users feel more *emotionally engaged* when interacting with an embodied VA [11].

Prior work shows that intelligent robots and agents that are anthropomorphic in shape communicate emotion through facial expressions, gestures, and postures [2]. On the other hand, non-humanoid robots can be designed with emotional expressions using robotic movements [10].

In this work, we explore various modalities of nonverbal emotional expressions to indicate a voice assistant's conversational states and evaluate their efficacy in promoting user engagement. We first analyze the common conversational states involved in Human-VA interaction. Then for each conversational state, we identify an emotion that it possibly activates based on a literature review [3] and a survey with 78 VA users. In terms of emotion representation design for VAs, we propose *facial expression*, *text box movement*, and *visual voice waveform* modalities. A within-subject experiment was conducted with 24 participants to compare the effectiveness of the three expressive designs. The results show that people are more emotionally engaged in the facial expression and text box movement modes than voice waveform mode. Meanwhile, the three design modes of VAs help maintain the same level of cognitive engagement.


















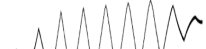





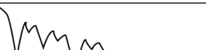
**Table 1:** The state-emotion mapping

<i>State</i>	<i>Description</i>	<i>Emotion</i>
Standby	Absence of a desired stimulus	Bored
Listening	Appearance of a desired stimulus	Curious
Receiving, Parsing	Uncertainty in stimulus processing	Confused
Intending	In readiness for converting the stimulus into an action	Eager
Acting	Execution of converting the stimulus into an action	Excited
Reporting	Success in achieving goal	Joyous
Interpreting, Reporting	Difficulty in achieving goal	Frustrated

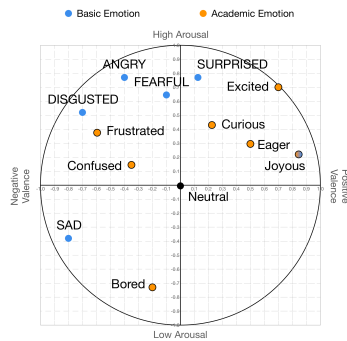
**State-Emotion Mapping**

Our goal is to design voice assistants (VAs) a set of emotional expressions for the purpose of indicating their conversational states and improving user engagement. To this end, we first analyze the conversational states involved in Human-VA interaction and further map emotions onto these states. The resulting state-emotion mapping will be used to drive Human-VA conversational interaction.

By adapting Brennan and Hulstee's conversational feedback model for a voice interface [4], we summarize a sequence of eight conversational states as summarized in the left side bar. Emotions were mapped onto these conversational states based on Breazeal's method [3], which relates emotions of a robot to its antecedent conditions. Some of the conversational states (*i.e.*, state 1, 2, 5, 8) can

	Facial Expression	Text Box Movement	Voice Waveform
<i>Idle</i>	 <p>Regular opening eyes; closed lips; raised lip corners</p>	 <p>Static</p>	 <p>Medium velocity; medium wavelength; medium amplitude; rounded, sinusoidal shapes; level trend</p>
<i>Bored</i>	 <p>Sleepy eyes; droopy lip corners; sometimes yawn;</p>	 <p>Bowing its body; falling asleep</p>	 <p>Slow velocity; long wavelength; low amplitude; short triangular shapes with long, flat plateaus</p>
<i>Curious</i>	 <p>Looking around; lips slightly apart;</p>	 <p>Craning its neck around to look</p>	 <p>Fast velocity; short wavelength; high amplitude; triangular shapes with fast rise; slightly upward trend</p>
<i>Confused</i>	 <p>Looking up and down; slightly pursed lip;</p>	 <p>Tilting its head to one side</p>	 <p>Medium velocity; medium wavelength; tangled shapes; slightly downward trend</p>
<i>Eager</i>	 <p>Blinking eyes; a nodding head; lips apart; raised lip corners</p>	 <p>Bobbing up and down slightly and quickly</p>	 <p>Fast velocity; short wavelength; high amplitude; rounded shapes; upward trend</p>
<i>Excited</i>	 <p>Big rounded opening eyes; lips widely apart; raise lip corners</p>	 <p>Jumping and dancing</p>	 <p>Greatly fast velocity; greatly short wavelength; greatly high amplitude; sharp peaks; upward trend</p>
<i>Joyous</i>	 <p>Smiling eyes; lips slightly apart; raised lip corners</p>	 <p>Swaying from left to right</p>	 <p>Slightly fast velocity; slightly short wavelength; slightly high amplitude; triangular shape with rounded corners; upward trend</p>
<i>Frustrated</i>	 <p>Frowning eyes; lips slightly apart; droopy lip corners</p>	 <p>Lowering its head slowly; rapid foot tapping; trembling</p>	 <p>Fast velocity; short wavelength; high amplitude; sharp peaks with rapid fall; downward trend</p>

**Table 2:** The design of emotion expression for a voice assistant, including facial expression, text box movement, and voice waveform. Each expression design includes a set of emotion illustrations.



**Figure 1:** The basic and academic emotions placed on Russell's pleasure-arousal grid, the dots with strokes represent the emotions used in our design.

be directly associated with these conditions. For example, an absence of the desired stimulus is associated with the standby state. The other conversational states (*i.e.*, state, 3, 4, 6, 7) are mapped onto emotions based on a survey with 78 VA users. In the survey, users were required to assign an emotion to each state from a set of 20 emotion terms collected from both basic emotions and academic emotions. Each state was labeled by the emotion term with the most selection. Table 1 summarizes the state-emotion mapping.

## Expressive Design of Emotions

We introduce three emotional expression designs: 1) facial expressions, 2) text box movements, and 3) voice waveforms, to indicate the aforementioned conversational states.

### Facial Expression

Facial expression is one of the most common methods to express nonverbal feedback during a conversation. To design facial expressions for VAs, we attempt to identify the primary facial elements such as eyes and then create emotional facial expressions using these elements.

A set of frequently used facial expressions elements were summarized based on the anatomy of well-received cartoon designs. These designs were shown to the 78 VA users in the sample survey and the ones with the most preferences were chosen as the primary facial elements for designing the emotional expressions. The design is inspired by prior research on emotion analysis [9], the investigation of emoji, and the observation of individual facial expression photos. Following Russell's pleasure-arousal grid (Figure 1), we propose a facial expression design by deforming the shape of the eyes and the mouth to encode the changes of valence along the x-axis and the changes of arousal along the y-axis (see Table 2, Column 1).

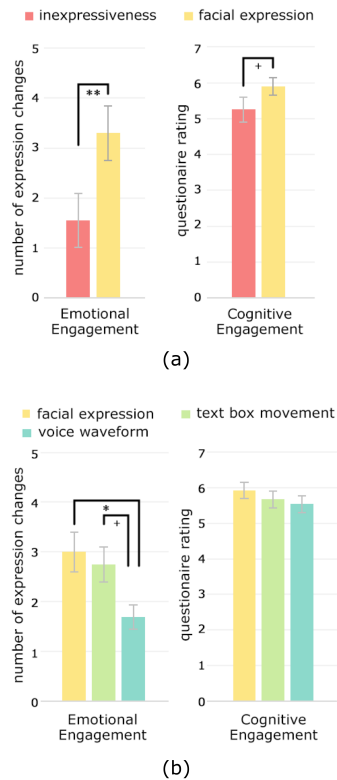
### Text Box Movement

Expressive body movements, which is a sequence of varied positions, constitutes an additional form of communication. Applying movement to icons in GUI will make the associated operations more intuitive and expressive [6]. We propose to use text boxes displaying textual responses of VAs as a medium of the embodiment to simulate body movements and imply emotions.

Our design (see Table 2 Column 2) were inspired by the existing research on affective body expressions [5] and kinetic GUI elements [6]. We derive a general principle of the composition of the bodily expression design, and project them onto the text boxes to communicate the current status of the VAs: postures are used to encode the valence dimension of an emotion while kinematics are used to encode the arousal dimension. The form of body squeezes when the emotion is negative and stretches when positive. From low to high arousal, kinematics increase.

### Voice Waveform

Waveform, as the most commonly used method for visualizing the voice, can also be used for expressing emotions. Existing study [7] suggests that most emotional vocal features are correlated to arousal, for example, high speech rate, intensity, and pitch implies high arousal. In our design (see Table 2, Column 3), we map these acoustic properties onto the visual attributes of a sine waveform: wave velocity encodes speech rate, amplitude shows vocal intensity, and wavelength encodes pitch. The resulting waveform is further mixed with a voice quality signal (*e.g.*, breathy, creaky, and whispery) to express emotion [8]. The trend of the waveform also used to illustrate the change of emotion. A rising trend indicates an emotional change from negative to positive.



**Figure 2:** Means and standard errors of the user engagement in our (a) pilot study and (b) user study (+:  $.05 < p < .1$ , \*:  $p < .05$ , \*\*:  $p < .01$ ). Emotional engagement is measured by number of users' expression changes while cognitive engagement is measured by user ratings on a 7-point Likert scale.

## User Study

In the pilot study with 20 participants (9 females), we found that a VA with an expressive face evokes significantly stronger user engagement than the one without any nonverbal expressive feedback, as shown in Figure 2(a). Therefore, we focused the user study on comparing the effectiveness of the three emotional expression designs for VAs.

### Method

In our within-subject experiment with 24 participants (15 females), three emotional expression designs were integrated to a VA as *facial expression* mode, *text body movement* mode, and *voice waveform* mode. The experiment consisted of three sessions, each of which involved one of the three modes. In each session, participants completed three interaction tasks using VAs (e.g., setting a reminder, making a reservation). To minimize learning effects, we counterbalanced the order of the three modes of the assistant.

We evaluated the effectiveness of emotional expression designs on user engagement in Human-VA interaction, including emotional engagement and cognitive engagement. To measure *emotional engagement*, two external proctors were instructed to record the number of facial expression changes that participants presented during the user study [14]. At the end of each session, we ask the participant to complete a user experience questionnaire about *cognitive engagement* using a 7-point Likert scale.

### Results and Findings

Repeated measures one-way ANOVA is applied to compare the user engagement of the expressive VAs. As shown in Figure 2(b), participants presented the most facial expression changes when interacting with the VAs with facial expression ( $M = 3.00$ ,  $SD = 1.93$ ), followed by those with text box movement ( $M = 2.75$ ,  $SD = 1.70$ ) and voice waveform ( $M = 1.70$ ,  $SD = 1.22$ ). Significant differences among user

emotional engagement are detected ( $F_{2,46} = 4.10$ ,  $p < 0.05$ ,  $\eta^2 = 0.11$ ). Tukey HSD post-hoc reveals significant differences between *facial expression* and *voice waveform* ( $p < 0.05$ ) and between *text box movement* and *voice waveform* ( $p = 0.08$ ). In terms of cognitive engagement, there is no significant difference among the three modes (*facial expression*:  $M = 5.92$ ,  $SD = 1.14$ ; *text box movement*:  $M = 5.67$ ,  $SD = 1.13$ ; *voice waveform*:  $M = 5.57$ ,  $SD = 1.16$ ). Participants showed different preferences when interacting with expressive VAs. “I think the changes of facial expressions are more understandable” (S2P16). “I like the waveform for daily use. It has less visual complexity but still conveys emotions” (S2P24).

In our study, we observed that participants smile more when the VA with facial expressions is showing a joyous face for achieving goals or blinking eyes with eager for requiring permission. Meanwhile, the VA with text box movements prompts more bodily entertainment in participants; they moved more when the text boxes are jumping, dancing, or swaying. Future work can explore modalities of emotional expression to induce different dimensions of user engagement, such as emotional engagement (e.g., laughter) and behavioral engagement (e.g., gaze).

We also found that the emotions of positive valence and high arousal help establish emotional connections between users and VAs. Participants presented more facial expression changes when the VAs with facial expressions or text box movements convey joy, eagerness, and excitement (see Figure 1).

## Conclusion

We propose three emotional expression designs, facial expression, text box movement, and voice waveform, to indicate a voice assistant's conversational states. Our study

results show the effectiveness of the emotional expression designs on user engagement in Human-VA interaction. Further studies exploring modalities of affective designs that accommodate to more scenarios will advance our understanding of the role of emotion in Human-VA relationship.

### Acknowledgements

Nan Cao is the corresponding author. This research was sponsored in part by the Fundamental Research Funds for the Central Universities in China No.22120180012.

### REFERENCES

1. Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 25–32.
2. Aryel Beck, Lola Cañamero, and Kim A Bard. 2010. Towards an affect space for robots to display emotional body language. In *RO-MAN*. IEEE, 464–469.
3. Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies* 59, 1 (2003), 119–155.
4. Susan E Brennan and Eric A Hulteen. 1995. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-based systems* 8, 2-3 (1995), 143–151.
5. Yu Chen, Xiaojuan Ma, Alfredo Cerezo, and Pearl Pu. 2014. Empatheticons: Designing emotion awareness tools for group recommenders. In *Proceedings of the International Conference on Human Computer Interaction*. ACM, 123–130.
6. Chris Harrison, Gary Hsieh, Karl DD Willis, Jodi Forlizzi, and Scott E Hudson. 2011. Kineticons: using iconographic motion in graphical user interface design. In *Proceedings of SIGCHI*. ACM, 1999–2008.
7. Patrik N Juslin and Petri Laukka. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin* 129, 5 (2003), 770.
8. Marko Luggner and Bin Yang. 2008. Psychological motivated multi-stage emotion classification exploiting voice quality features. In *Speech Recognition*. InTech.
9. Xiaojuan Ma, Jodi Forlizzi, and Steven Dow. 2012. Guidelines for depicting emotions in storyboard scenarios. In *International Design and Emotion Conference*.
10. Jekaterina Novikova and Leon Watts. 2014. A design model of emotional body expressions in non-humanoid robots. In *Proceedings of the 2nd International Conference on Human-Agent Interaction*. ACM, 353–360.
11. Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence* 19, 3-4 (2005), 267–285.
12. Victoria L Rubin, Yimin Chen, and Lynne Marie Thorimbert. 2010. Artificially intelligent conversational agents in libraries. *Library Hi Tech* 28, 4 (2010), 496–522.
13. Daniel Szafrir and Bilge Mutlu. 2012. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of SIGCHI*. ACM, 11–20.
14. Qianli Xu, Liyuan Li, and Gang Wang. 2013. Designing engagement-aware agents for multiparty conversations. In *Proceedings of SIGCHI*. ACM, 2233–2242.