Expressive Auditory Gestures in a Voice-Based Pedagogical Agent

Jessy Ceha
School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada
jceha@uwaterloo.ca

Edith Law
School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada
edith.law@uwaterloo.ca

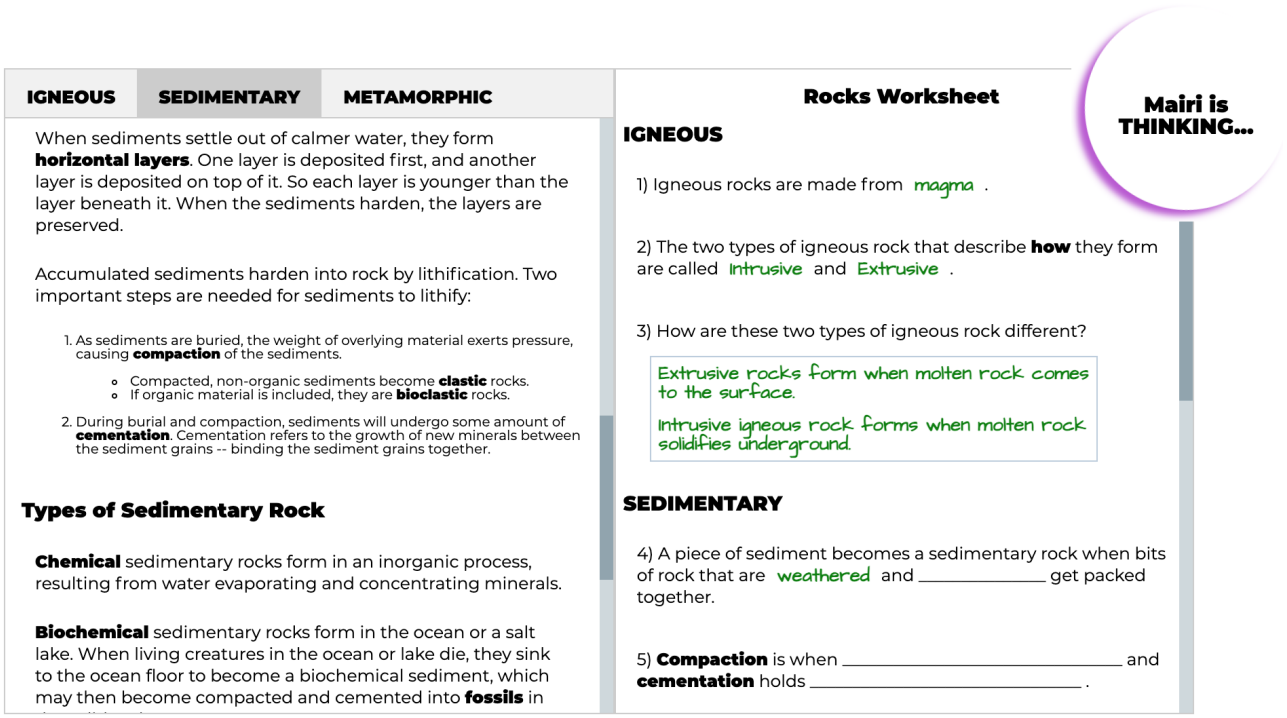

Figure 1: Platform on which participants taught the voice-based agent Mairi about various rock types by using the information provided in the left panel and guiding the agent in filling out the worksheet in the right panel.

ABSTRACT

In this paper, we explore how expressive auditory gestures added to the speech of a pedagogical agent influence the human-agent relationship and learning outcomes. In a between-subjects experiment, 41 participants assumed the role of a tutor to teach a voice-based agent. The agent used either: expressive interjections (e.g.,"yay", "hmm", "oh"), brief expressive musical executions, or no auditory gestures at all (control condition), throughout the interaction. Overall, the results indicate that both gestures can positively affect the interaction, but in particular, interjections can significantly increase feelings of emotional rapport with the agent and enhance motivation in learners. The implications of our findings are discussed as our work adds to the understanding of conversational agent design and can be useful for education as well as other domains in which dialogue systems are used.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in interaction design**.

KEYWORDS

education, voice, rapport, motivation, agent, interjections, music

ACM Reference Format:
Jessy Ceha and Edith Law. 2022. Expressive Auditory Gestures in a Voice-Based Pedagogical Agent. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3491102.3517599

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517599

1 INTRODUCTION

Conversational agents — systems that imitate natural language discourse — can take on various forms (i.e., virtual, physical, embodied, non-embodied) and ways of expression (i.e., verbal, non-verbal, etc.). Among the different roles conversational agents can play (e.g., assistant, companion), they have become a widely used tool in learning environments. Pedagogical agents (i.e., conversational agents for education) have been designed for a number of purposes, such as tutoring (e.g., [65]), language learning (e.g., [38]), and promotion of skills such as metacognition (e.g., [49]). Historically, learning technologies have focused heavily on learners' cognitive needs. However, researchers are becoming increasingly aware of the importance of affective factors (emotions, feelings, and moods) in educational technologies, and there is evidence to suggest that the manner in which pedagogical agents converse with learners as well as the relationship that develops between them, can enhance various learning outcomes (e.g., motivation and recall) [6].

Prior work on conversational agents across various domains suggests that expression of internal states has a positive influence on the human-agent relationship, e.g., modifications of voice pitch was found to improve interaction quality with a social robot [43], and body gestures and speech content increased perceived personality of a virtual agent [42]. Many studies have focused on expression through facial cues, and bodily postures or gestures. A lesser studied form of expression in conversational agents is the addition of brief auditory gestures. Similar to visual cues, auditory gestures are able to convey information about internal state [8]. In this work, we look specifically at two types of auditory gesture at the word-level, namely interjections and brief musical executions. In human-human dialogue, vocal gestures known as *interjections* are used to communicate the speaker's emotions and attitudes. They can include both speech and non-speech utterances, e.g., "yay", "hmm", "argh". In human-agent interaction, agents can express themselves through interjections but also language-independent sounds. One expressive type of sound is *music*, as music and the human voice have a number of similarities when it comes to expressing emotions and intent [27].

We describe in this paper an experiment in which we systematically manipulate the addition of expressive auditory gestures to the speech of a voice-based agent. We look specifically at applying these gestures in an educational context in which the agent takes on the role of a novice and the user a tutor – in order to investigate the role of auditory gestures in learning-by-teaching. We compare two types of gestures: *interjections* and their semantic-free counterpart, *music*. Both types allow for emotion, cognitive, and intent expression, but musical sounds have no language dependence. These word-level cues contrast prior work that has looked at such things as utterance-level prosody manipulations for example, and our results generate new insights on how users perceive an agent that adds expressive auditory gestures to communicate its internal state, and what effects these gestures have on the interaction. The findings add to our understanding of pedagogical agent, and more generally conversational agent, design: how are interjections and music perceived and what impacts do they have on the interaction as well as learning outcomes in an educational context.

In the following section we provide a review of the related work in this area. Thereafter, we introduce our study design and procedure, followed by the results and a discussion with suggestions for future work.

2 RELATED WORK

2.1 Conversational Agents in Education

Conversational agents are systems that mimic human natural language discourse and engage users in conversation. The natural language can be communicated through voice or text, and the agents can take on various forms and ways of expression, commonly being categorized as chatbots, virtual avatars, or physical robots. Conversational agents are deployed in numerous domains including healthcare, entertainment, customer service, and education [33, 62]. For learning, educational technologies started largely with intelligent tutoring systems — designed to provide such things as timely feedback and support active engagement. However, certain instructional strategies are based in dialogue: asking-deep reasoning questions, self-explanations, collaborative interactions, and fostering common ground [45], and conversational agents are able to provide this verbal discourse to learners.

2.1.1 Teachable Agents. Pedagogical agents are made for learners of different ages and can be distinguished along various dimensions including topic: such as science (e.g., [65]) or language learning (e.g., [39]), form & appearance: chatbot, virtual, or physical, as well as role: either as teacher agents, co-learner agents, or teachable agents [31]. In this work, we focus on teachable agents — systems that learn with, and are taught by, the learner, based on the concept of learning-by-teaching; a widely studied and practiced technique within the education domain, as learning-by-teaching can be a more enriching experience than learning by oneself [16]. In contrast to a teacher agent, teachable agents struggle with the material and make errors, which prompts the human learner to pay more attention, reflect on misconceptions, and elaborate on explanations [50]. Teachable agents have been found to be effective in computer-based learning environments. For example, Chase, Chin, Oppezzo, and Schwartz [11] demonstrated that students put in more effort (i.e., spending more time on learning activities and learning more) for a virtual teachable agent than they did for themselves, and interacted socially with them — attributing mental states and responsibilities to them.

2.1.2 Learning Outcomes. The overarching goal of pedagogical agents are to elicit learning outcomes. Learning outcomes can be considered as changes in the following three areas: cognitive, skill-based, and affective [29]. *Cognitive outcomes* are focused around the constructs of declarative knowledge, knowledge organization, and cognitive strategies; *Skill-based outcomes* are concentrated on performance-type measures such as speed and fluidity of performance; and *Affective outcomes* include both attitudinal and motivational outcomes [30].

Motivation — to be moved to do something [51] — is an especially important affective outcome as it can influence what, when, and how we learn [57]. People vary in their levels of motivation (i.e., how much) and in their orientation of motivation (i.e., the type of motivation). Deci and Ryan [14] distinguish between different types of motivation: intrinsic motivation (doing something

because of interest or enjoyment) and extrinsic motivation (doing something because of external prods, pressures, or rewards), and argue for the importance of both types of motivation for successful learning [51]. Research in this area has investigated various ways of promoting motivation in learners. Saerbeck, Schut, Bartneck, and Janse [52] had children interact with one of two types of robotic language tutor: one which engaged the child in a social dialogue or one which was neutral and focused on knowledge transfer. The social robot employed strategies such as saying "we" instead of "you", making motivational statements, and using non-verbal and verbal gestures. The researchers found that the social-supportive robot had a positive effect on the learning performance of participants, and that intrinsic and task motivation were significantly higher in the social condition. Liew, Zin, and Sahari [34] found that a virtual agent's enthusiasm – conveyed through the tone of voice, constant smiling, a high level of animated movement and head-nodding during speaking, as well as enthusiastic remarks – significantly enhanced university students' emotion, intrinsic motivation, affective perceptions, and cognitive outcome.

Teachable agents present a unique opportunity for exploring motivation enhancement as evidence suggests that when learners taking on the role of the tutor, feel more responsible or have a better relationship with their agent, they are more motivated to put in effort to teach their agent, and as a result also learn more [32, 46].

Overall, prior work on pedagogical agents indicates that motivation in learners can be enhanced by the manner in which agents converse with them, as well as the relationship that develops between learner and agent. In order to build and strengthen relationships, research on conversational agents in education and beyond, suggests the importance of expressing internal states. For example, the emotional coloring of an utterance (i.e., the activation (active/passive), evaluation (positive/negative), and power (dominant/submissive) has been found to enhance feelings of rapport with a voice-based agent [1]; variations in voice pitch can result in increased ratings of interaction quality with a social robot [43]; body gestures and speech content increase perceived personality of a virtual agent [42]; and facial expressions in an agent were shown to positively influence learning outcomes [7]. Much of this prior work has focused on expression through visual cues (i.e., gestural or facial), or utterance-level modifications of speech. Instead, we look at two types of auditory gesture: interjections and brief musical executions.

2.2 Expressive Auditory Gestures

2.2.1 Interjections. Interjections are parts of speech, a word or phrase, that express the internal state of the speaker — their emotions and attitudes. Expressive interjections can express a reactionary feeling or emotion such as surprise, delight, fear, disgust (e.g., "ah", "aww", "blah", "bother", "eww", "good grief", "oh", "ugh"), and convey feelings that result from what one comes to know or understand (e.g., "aha", "yay", "gee", "huh", "oh", "hmm", "golly"). Although interjections are common in human dialogue, and are suggested to strengthen relational bonds between humans and agents [53], only recently have researchers started focusing on the addition of interjections to the speech of conversational agents.

Cohn, Chen, and Yu [13] introduced interjections (e.g., "Wow") and fillers (words used by the speaker to manage the dialogue; e.g., "um", "uh", "like") to the voice-based Amazon Alexa agent. The researchers found that interjections and fillers separately improved overall user ratings (n=5,527), with a further increase observed if they were used simultaneously. An additional perception study supported the findings with interjections leading to higher social ratings — especially in engagement, naturalness, expressiveness, and likeability. Hu et al. [23] present preliminary findings of their emotionally aware voice-based conversational agent called HUE. Following sentiment analysis of the human speaker, HUE would use an interjection that reciprocated the same emotion (e.g., "wow", "haha"). 75 participants observed HUE interact with people in various scenarios and rated its perceived emotional intelligence significantly higher when HUE responded to emotion with interjections than not.

2.2.2 Music. In human-agent interaction, agents can express themselves through interjections but also vocalizations and sounds that have no semantic content or language dependence (also known as semantic-free utterances, e.g., [67]). Such sounds are a unique mode of communication for conversational agents and can include gibberish, non-linguistic sounds (e.g., beeps, squeaks, clicks), musical expressions (e.g., musical tones or instrument sounds), and paralinguistic vocalizations (e.g., moans and laughter). Some work has looked for example at adding such sounds to the synthetic speech of an agent to enrich agent characterization. Aylett, Vazquez-Alvarez, and Butkute [5] conducted a preliminary study exploring how the addition of various sounds and vocalizations impacted perceived personality of a social robot. They created five semantic-free utterances to communicate: agreement, disagreement, curiosity, sadness, and amusement. The results indicated that the utterances made the voice seem more extrovert, but participants also noted feeling that the additional sounds were disconnected from the speech. As the authors point out, designing semantic-free and language-independent vocalizations and sounds is complex and there exists no systematic process for generating them to convey specific affective states.

However, of the different types of semantic-free utterances, a number of similarities have been found between musical expressions and the human voice in conveying emotions and intent [27]. Moreover, music has been found to elicit and influence emotional states [28] as well as enhance cognitive abilities such as learning and memory (e.g., [37, 54]; for a review, see [17]). In human-computer interaction research, music has been studied as a means of debugging software [63] and communicating information to blind or visually impaired users [3]. Researchers have also started looking at using music to convey the internal state of agents. Jee and colleagues [25, 26] and Jee, Jeong, Kim, and Kobayashi [24] across a number of studies, created various musical expressions by analyzing the sounds of the robots R2-D2 and Wall-E. They designed sounds to convey particular intentions (affirmation, denial, encouragement, introduction, question) and emotions (happy, sad, shy, fear, and dislike) in an English teaching robot, and found high recognition rates of the intended internal states.

In summary, prior work has investigated the addition of musical expressions and other sounds to conversational agent speech,

looking at how they should be designed, whether they can be perceived as intended, and what the effects are on perception of the agent, however there is little research on adding interjections or musical expressions to pedagogical agent speech and the impacts on relationship building and learning outcomes. As the relationship between human and agent has been found to enhance motivation in learners, and expressions of internal state are suggested to bolster this relationship, with this study we set out to explore what impacts the addition of brief expressive auditory gestures to the synthetic speech of a voice-based agent have on the human-agent relationship as well as affective (i.e., motivational) and cognitive (i.e., recall) outcomes, in a learning-by-teaching scenario with a teachable agent. Given the results of prior work, we hypothesized that separately, interjections and musical executions, would result in higher ratings of rapport and interaction quality, and that there would be an increase in the affective learning outcome – motivation – with increased rapport. In the context of learning-by-teaching with teachable agents, the increased motivation would lead to improvement in the cognitive learning outcome – recall.

3 STUDY DESIGN

3.1 System

Task. We built an application which allowed participants to teach a voice-based agent about various rock types. The interface provided to participants (Figure 1) contained information about rocks, which was adapted from a Lumen Learning course at the Geology 101 level. Lumen Learning was used as it is an open-educational resource developed for university/college students – the main demographic of our participant sample. Participants also saw a 'Rocks Worksheet' on the right-hand side of the interface, which the agent needed to fill in; it had some knowledge of the topic but required help from participants to confirm or correct its prior knowledge and learn new knowledge. The information on rocks in the left panel of the interface was divided into three tabs corresponding to rock type: Igneous, Sedimentary, and Metamorphic; and the Rocks Worksheet similarly contained questions on each type, with both fill-in-the-blank, as well as longer-form questions. Participants used the interface to read through the information and communicate with the agent by pressing on the button in the top-right corner to record and send messages. The agent responded to the participant and filled in the worksheet (green text in Figure 1) as it was taught.

Wizard. The agent's responses were controlled by a human operator (first-author) using the Wizard of Oz technique [40], with a set of pre-defined statements, to reduce the system response time and maintain a similar conversation across participants (see Figure 2). The system was built using WebSockets and Django Channels to allow for real-time communication between participant and agent (Wizard). All pre-defined statements were buttons which the Wizard could click on when appropriate. The left-hand side of the interface was designed to mirror the left-hand side of the participant's interface, with tabs along the top to organize statements based on rock type (Igneous, Sedimentary, and Metamorphic). The Wizard's interface had an extra tab 'Intro' which contained statements to initiate a short dialogue between the agent and participant whereby the agent introduced itself, asked for the participant's name, briefly repeated the task, and asked which rock type/set of questions the participant wanted to start with. The buttons/statements in each rock type tab were designed to follow a script revolving around each question on the Rocks Worksheet. Grey and orange coloured buttons contained speech, while red buttons comprised the agent's intended answers for the questions on the Worksheet. Orange buttons indicated that the speech contained an additional expressive auditory gesture. The Wizard could also type out a different answer for the Worksheet, e.g., "Enter worksheet answer Q4a:", if the participant taught information that deviated from the pre-defined answer in the Wizard's interface. Similarly, on the right-hand side of the interface, the Wizard could view the chat log between themselves ("Wizard") and the participant (e.g., "p1"), and if necessary use the dialog box to type responses that were not pre-set. The middle section of the interface contained statements that were *Common* across the interaction, i.e., not constrained to a type of rock. Lastly, the blue button ends the interaction, disabling the ability for the participant to send any further voice messages to the agent. In case the participant asked an off-topic question, the agent would respond that they did not understand and/or reverted the conversation back on-topic.

Speech Recognition. To speak to the agent, the participant clicked on a button in their interface. The Web Speech API, a JavaScript Web Speech API Specification, was used to access the participant's browser audio stream and convert it to text. The text was then sent to the Wizard and stored for later analysis. The Wizard could also hear the participant through the online conferencing tool in which the study was being held, so as to increase accuracy of understanding the participant.

Speech Synthesis. After receiving a participant's message, the Wizard selected a response from the set of pre-defined statements on their own interface. For text-to-speech of the agent, CereProc was used (https://www.cereproc.com/), renowned for synthetic voices retaining naturalness and character — "The CereVoice Engine SDK is the first free, commercial-grade, real-time speech synthesis system for academic research. It is fast, stable, and highly configurable, and is well suited to research into text-to-speech and dialogue applications." CereProc voices allow for emotional synthesis control and each voice comes with vocal gestures such as laughs, coughs, and expressive interjections, e.g., "hmm", "ah", "yeah", "oh", etc. The voice chosen was Mairi - a child's voice with a Scottish-English accent. This voice was selected as, of the voices available, it was determined by the authors through pilot testing, to pronounce the words used in our conversational scenario most clearly and comprehensible for participants.

3.2 Conditions

We used a between-subjects experimental design, with participants being randomly placed into one of three conditions: (1) Interjections, (2) Music, or (3) Control (no added auditory gestures).

The agent made both correct and incorrect statements about the topic being taught throughout the interaction. These were determined by the script that was designed for each question on the Rocks Worksheet, and was therefore consistent across all participants. The statements and dialogue flow were developed through iterative pilot studies. Both correct and incorrect statements lead to moments where the agent replied with auditory gestures of

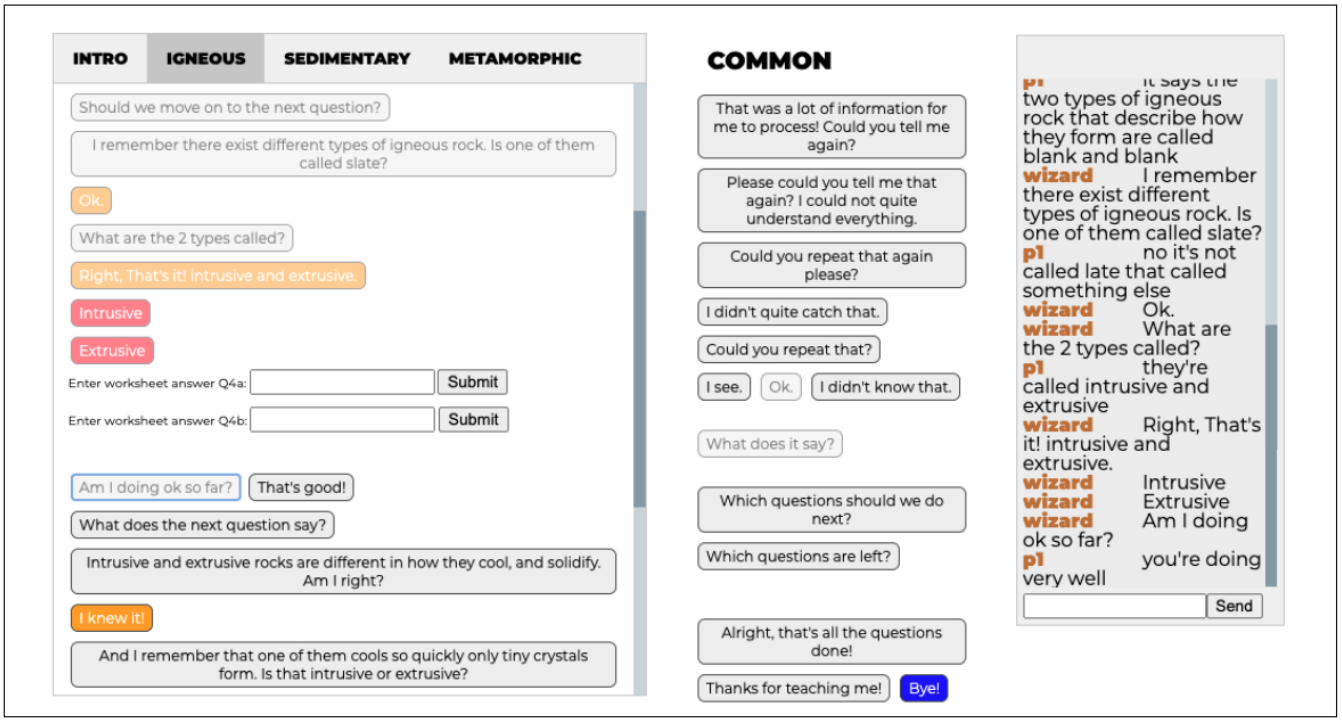

Figure 2: Wizard interface during experiment. This is an example interaction between the Wizard and a participant, p1. The Wizard could view the chat log (on the right) and selected the appropriate pre-defined statement for the agent to say from the options on the left (i.e., statements organized by rock type and question on the Rocks Worksheet) or in the center (i.e., common statements independent of rock type).

positive or negative valence, respectively. Valence ranges from pleasant (positive) to unpleasant (negative) – examples of negative valence include sadness and fear, whereas happiness is an example of positive valence. There was an equal amount of both through the conversation: eight moments of positive, and eight moments of negative. The moments were selected by a native English speaker (first author) based on the fit within the conversational context. For each participant, the exact gesture was randomly selected with a pseudo-random number generator from the set of either positive or negative ones, so as to increase variation.

A total of fourteen positive and negative **interjections**, provided by the CereProc voice, were used in the study: seven with positive valence ("sigh", "hmm", "hmmm", "ah", "oh", "yay", "yeah") and seven with negative valence ("sigh", "ah", "oh", "ugh", "argh", "arr", "doh"). The **musical sounds** were taken from the validated set of auditory stimuli: The Musical Emotional Bursts (MEB) dataset [48]. It consists of 80 brief musical executions expressing positive and negative valence. The dataset contains both improvisations and imitations of emotional expressions made by violinists and clarinetists. We chose to use the improvisations of the clarinetists as they were shown by Paquette et al. [48] to have a high recognition accuracy for both the positive and negative valences, resulting in a total of fourteen musical executions: seven positive and seven negative.

Each speech utterance that followed an auditory gesture (i.e., interjection or music) was also matched in valence (positive or negative) to the gesture, using the synthesis control provided by the CereProc voice. For example, if the gesture was positive, then the utterance following it, i.e., corresponding to it, was also synthesized with positive valence. This was done so that the agent expressed valence in all conditions including control, with the only difference being the added expressive auditory gestures in the two experimental conditions. Adjusting the gesture and speech utterance to match is also important as prior work indicates users can feel a disconnect between speech and sound if not matched appropriately (e.g., [5]). No specific personality of the agent was designed, however the agent made three off-task statements during the conversation: indicating enjoyment of the task, asking how they're doing so far on the task, and telling a rock joke, as well as asking some additional on-topic questions about rocks. These off-task moments were included as prior work indicates they can lead to more positive experiences in pedagogical agent interactions (e.g., [22]).

3.3 Procedure

The study began by a researcher and participant connecting via an online conferencing tool. From here, participants were given an information letter and consent form, and then asked to fill in a pre-study questionnaire that included demographics information (e.g., age, gender, experience with conversational agents, etc.) and

AGENT	What's the first question about metamorphic rocks?
USER	It says metamorphic rocks are made by blank.
AGENT	So metamorphic rocks are the ones that form from intense heat and pressure. Am I right?
USER	That's exactly right. You got it.
AGENT	***Yay***! There we go. [positive valence]
AGENT	Shall we move on to the next question?
USER	Yes good idea. The next question asks: what is foliation in metamorphic rocks caused by?
AGENT	What causes foliation? Well first of all, what does foliation mean? Is that when there are fossils in the rock?
USER	No it's not fossils. Foliation are the flat layers that form as the rock is squeezed by pressure.
AGENT	***Oh***... So it's not fossils, but layers in the rock. [negative valence]

Table 1: Example conversation transcription between a participant and the agent in the Interjections condition.

a quiz to test their knowledge on the topic to be discussed during the interaction with the agent. The quiz included 14 questions (5 for each rock type - 1 question covers both metamorphic and sedimentary types) and were adapted from the Lumen Learning course. Three multiple choice answers were provided for each question. Some quiz questions were covered in the interaction with the agent directly, while others were in the articles but not covered specifically by the agent's questions.

Participants were then introduced to the interface (Figure 1) on which they would complete the task. After signing in to the interface, participants were given 3 minutes to read through the articles and questions on the 'Rocks Worksheet'. Following this, the agent introduced themselves to the participant, asked them their name, and explained their task again briefly. Participants had this preliminary dialogue with the agent to reduce novelty effects before the actual experiment began. Once participants communicated to the agent they were ready to begin, the agent asked which worksheet questions on one of the three rock types (Sedimentary, Metamorphic, Igneous) the participant wanted to start with. This was done to provide participants with an opportunity to partially guide the interaction and to increase variation of the order in which the questions were covered. The interaction with the agent lasted around 20-30 minutes. Following the interaction, participants filled out a number of post-study questionnaires. In total, all questionnaires took approximately 15-20 minutes.

3.4 Measures

3.4.1 Human-Agent Relationship. To measure participants' perceived relationship with the agent, we adapted the rapport instrument used by other researchers [20, 44] - covering two rapport dimensions, Understanding (a sense of mutual understanding) and Emotional (a sense of emotional connection), and extended it with questions from [9] on Quality of Interaction. The 15 questions were presented with a five-point Likert scale from "strongly disagree" (1) to "strongly agree" (5), balanced for positive and negative responses. Pick-a-Mood, a cartoon-based pictorial instrument, was used to measure the perceived mood of the agent [15], and finally, participants were given a number of questions relating to experience of the teaching task and the agent as a student: "How much did you like teaching [the agent]?", "Do you think you were good at teaching [the agent]?", and "Do you think [the agent] was a good student?". They rated their agreement with each question by selecting one of the following: not at all, a little, quite a lot, and very much, and were also provided the opportunity to give free-form answers.

3.4.2 Learning Outcomes. To evaluate cognitive learning outcomes, participants were asked to take the same quiz as prior to the interaction and we compared changes in quiz score pre- and post-interaction to measure recall of the material. To investigate the affective learning outcome motivation, we used the *Academic Motivation Scale* (AMS) [61]. The wording of questions varies slightly from the original to fit the study. The AMS provides overall scores for *intrinsic motivation* (IM; actions motivated by the pleasure and satisfaction from the process of engaging in an activity), *extrinsic motivation* (EM; actions motivated by attaining a goal separate from the process of engaging in an activity), and *amotivation* (AM; the absence of motivation which can co-occur with feelings of low competence). IM and EM can be further distinguished into more specific motives [64]:

- *IM - to know* describes actions performed for the pleasure and satisfaction derived from learning, exploring, or trying to understand something new
- *IM - toward accomplishment* relates to engaging in actions for the pleasure and satisfaction experienced when trying to achieve something new or beyond one's limits
- *IM - to experience stimulation* describes the motivation related to the experiencing of pleasurable sensations
- *EM - externally regulated* indicates the behaviour is motivated by reasons external to the task at hand, i.e., payment or rewards
- *EM - introjected* refers to actions motivated by pressure an individual puts on themselves
- *EM - identified* describes behaviour that is motivated by the view that participation is important for personal growth

3.4.3 Cognitive Workload. Based on cognitive load theory, some argue that pedagogical agents can impose extraneous cognitive load and be detrimental to learning outcomes (e.g., [12]). As our study involves dialogue with the addition of expressions that hold meaning, we wanted to investigate the impact on learners' cognitive load as well. The workload profile (WP; [60]) was used to investigate

subjective cognitive workload and had participants rate the proportion of attentional resources used on the following dimensions on a scale of 0 to 100:

- *perceptual/central processing*, i.e., "How much attention was required for activities like remembering, problem-solving, decision-making, perceiving (detecting, recognizing and identifying objects)?"
- *response selection and execution*, i.e., "How much attention was required for selecting the proper response channel (manual - keyboard/mouse, or speech - voice) and its execution?"
- *spatial and verbal processing*, i.e., "How much attention was required for spatial processing (spatially pay attention around you)?" & "How much attention was required for verbal material (e.g., reading, processing linguistic material, listening to verbal conversations)?"
- *visual and auditory processing*, i.e., "How much attention was required for executing the task based on the information visually received (eyes)?" & "How much attention was required for executing the task based on the information auditorily received (ears)?"
- *manual and speech output*, i.e., "How much attention was required for manually responding to the task (e.g., keyboard/mouse usage)?" & "How much attention was required for producing the speech response (e.g., engaging in a conversation, talk, answering questions)?"

The scores are then summed and for comparison, this sum is averaged.

3.5 Participants

The study received ethics clearance and 41 participants were recruited through mailing-lists and participant pools at a research-based institution in Canada. One participant did not consent to the use of their data, another did not comply with survey instructions, and another assumed from the beginning of the interaction that the study used the Wizard-of-Oz technique. Results reported are therefore based on 38 participants (21 women, 17 men; age range = 18-52 years; median = 24 years, SD = 5.75; 2 participants did not provide their age). All participants were volunteers and received a $15 gift card. Both native (61%) and non-native (39%) English speakers participated in the study, and varied in whether their post-secondary education related to STEM fields (84% STEM-related, 16% not), their highest completed or current degree (58% Bachelor's, 32% Master's, 8% Doctorate, 2% College credit), and their ethnicity (61% Asian, 18% White, 3% Aboriginal or Indigenous, 3% Asian and Native Hawaiian or other Pacific Islander, 3% Black or African American, 3% Asian and Hispanic, Latino or Spanish origin, 3% Middle Eastern or North African, 3% preferred not to disclose, and 3% self-described as Sikh Punjabi). Additionally, when asked to rate their level of interest and experience with conversational agents on a 7-point Likert scale from 1 (not at all/never) to 7 (very interested/a lot), 8% were not at all or had very little interest in agents, 47% were moderately interested, and 45% were highly interested; 26% had very little experience with agents, 42% had a moderate amount of experience, and 32% were highly experienced with conversational agents.

4 RESULTS

We collected both qualitative and quantitative data from each participant. For the numerous measures the following analyses were carried out: ANOVA, Kruskal-Wallis, linear regression models and cumulative link model (CLM; models the cumulative probabilities of discrete ordinal categories [2, 41]), with condition and demographics (gender, native language, pre-interaction quiz score, etc.) as the independent factors, and Pick-a-Mood, Understanding and Emotional Rapport, Quality of Interaction, AMS, WP, and change in pre to post quiz score, as the dependent factors.

4.1 Human-Agent Relationship

One-way ANOVA showed no significant difference between conditions on the measures of Quality of Interaction ($F(2, 35) = 1.54, p = .23$) and Understanding Rapport ($F(2, 35) = 1.63, p = .21$). Results of the free-form questions, analyzed with CLM, similarly showed no significant differences between conditions, and neither did the Pick-a-Mood pictorial self-report scale for the agent's mood and personality. However, condition was found to have an effect on the Emotional Rapport dimension ($F(2, 35) = 3.34, p = .05$), with Tukey's HSD showing that participants in the Interjections condition ($M = 4.48, SD = 0.43$) on average rated feeling significantly more Emotional Rapport with the agent than those in the Control condition ($M = 3.98, SD = 0.63$), at $p = .04$.

Participants' answers to the free-form post-interaction questions suggest that both the Interjections (I) and Music (M) gestures were perceived as intended, e.g., *"She seemed embarrassed or proud of herself at times. She expressed these emotions through verbal noises such as an exclamation when she would get the answer right or wrong"* (I09); *"Cheerful and interested. She showed this by exaggerated "no's" when an answer was incorrect and excited when she got an answer correct"* (I13), and *"[the agent] expressed "happiness" with a happy music and sadness with a "sulky music" " (M10)*; *"[the agent] was very expressive ... there was happy or sad music as well whenever it tried to reciprocate its emotion"* (M13).

The majority of participants in the Interjections and Music conditions stated they enjoyed the experience and expressed perceptions that, *"[the agent] was very personable"* (I07), *"she was engaged"* (I09), *"I liked the smooth interaction and the flow of the conversation"* (M01), *"it was interactive and engaging"* (M04), and *"the jokes and songs were fun and interesting"* (M09). Overall, most participants enjoyed the addition of auditory gestures during the task, but for a few participants the gestures were considered *"a little scary"* (M05) and *"a bit creepy"* (I01). M14 explained, *"I would leave out the music. It takes away from the flow of the conversation"*. Similar to work that found children distinguish between 'creepy' sounds that express intent and non-threatening sounds that are spontaneous [68] – our results suggest analogous sentiments in adults are possible. These statements also further support the importance of adjusting the gesture and speech to match, so as to maintain the flow of dialogue and enhance the experience.

4.2 Learning Outcomes

In terms of recall of material (i.e., cognitive learning outcome), quiz scores pre-interaction started relatively high in all conditions (out of 13) Interjections: $M = 7.07, SD = 2.16$; Music: $M = 8.08, SD = 1.50$;

	Control (n=11)	Music (n=13)	Interjections (n=14)	
Age (years)	M=23.91±3.14	M=24.08±1.88	M=26.85±8.91	$F(2, 33) = 1.02, p = .37$
Gender	5man, 6woman	5man, 8woman	7man, 7woman	$\chi^2(2, N = 38) = 0.37, p = .83$
Native English	9yes, 2no	6yes, 7no	8yes, 6no	$\chi^2(2, N = 38) = 3.28, p = .19$
Pre-quiz Score	M=7.91±1.22	M=8.08±1.50	M=7.07±2.16	$F(2, 35) = 1.33, p = .28$

Table 2: Demographic and condition data of study participants.

Control: $M = 7.91, SD = 1.22$, and on average increased post-interaction. Condition had no significant effect on change in quiz score from pre- to post-interaction, $F(2, 35) = 0.4, p = .67$. To investigate motivation (i.e., affective learning outcome) we used the AMS questionnaire which provided overall scores of intrinsic, extrinsic, and a-motivation, with intrinsic and extrinsic being further distinguished into more specific motives. Analysis of each overall and subscale score was done using one-way ANOVA followed by Tukey's HSD.

Extrinsic Motivation (EM). In terms of ratings of being extrinsically motivated to make the effort to teach the agent, there was a significant difference between conditions ($F(2, 35) = 6.19, p = .005$), with participants in the Interjections condition ($M = 4.26$, $SD = 1.29$) reporting on average significantly more extrinsic motivation than participants in the Control condition ($M = 2.82, SD = 1.08$), at $p = .004$. At a lower-level, while no significant differences were found in the *EM - externally regulated* or *EM - introjected* motives, there was a significant difference in the amount of *EM - identified* (behaviour motivated by the view that participation is important for personal growth) reported between conditions ($F(2, 35) = 9.83, p < .001$), with participants in both conditions: Interjections ($M = 5.50, SD = 0.96$), at $p < .001$, and Music ($M = 4.67, SD = 1.12$), at $p = .03$, rating their motivation in the task as *EM - identified* more highly than participants in the Control condition ($M = 3.39, SD = 1.47$). However, a linear model indicated that a higher pre-quiz score reduced the amount of reported *EM - identified* in both the Interjections ($\beta = -0.90, t(32) = -2.91, p = .007$) and Music ($\beta = -1.01, t(32) = -2.92, p = .006$) conditions compared to Control.

Intrinsic Motivation (IM). With regards to intrinsic motivation, condition was found to have a significant effect on overall intrinsic motivation ($F(2, 35) = 8.44, p = .001$), with participants in the Interjections condition ($M = 5.84, SD = 1.02$) feeling more intrinsically motivated than participants in the Control condition ($M = 3.84, SD = 1.34$), at $p < .001$, and some evidence of participants in the Music condition ($M = 5.03, SD = 1.29$) feeling more intrinsically motivated compared to Control as well, at $p = .06$. We also found condition to have a significant effect in each of the sub-motives of intrinsic motivation:

- *IM - toward accomplishment*: engaging in actions for the pleasure and satisfaction experienced when trying to achieve something new or beyond one's limits ($F(2, 35) = 5.08, p = .01$). Participants in the Interjections condition ($M = 6.07$, $SD = 0.94$) reported feeling significantly more *IM - toward accomplishment* than those in the Control condition ($M = 4.23, SD = 1.94$), at $p = .009$.
- *IM - to know*: actions performed for the pleasure and satisfaction derived from learning, exploring, or trying to understand something new ($F(2, 35) = 6.48, p = .004$). Participants in the Interjections condition ($M = 6.09, SD = 0.87$) reported feeling significantly more *IM - to know* than those in the Control condition ($M = 4.43, SD = 1.37$), at $p = .003$.
- *IM - to experience stimulation*: motivation related to the experiencing of pleasurable sensations ($F(2, 35) = 8.50, p = .001$). Participants in both the Interjections ($M = 5.36, SD = 1.50$) and Music ($M = 4.51, SD = 1.60$) conditions reported feeling significantly more *IM - to experience stimulation* than those in the Control condition ($M = 2.79, SD = 1.59$), at $p < .001$ and $p = .03$, respectively.

To investigate further the variables that impact the intrinsic motivation sub-motives, different linear models were used for each subcategory. Through step-wise selection, the resulting models show that participants' prior knowledge of the topic (i.e., higher pre-interaction quiz scores) significantly influenced feelings of *IM - to know* with participants in the Interjections ($\beta = -0.68, t(29) = -2.09, p = .05$) and Music ($\beta = -0.85, t(29) = -2.36, p = .03$) conditions reporting lower feelings of *IM - to know* than in the Control condition.

A-motivation. Lastly, one-way ANOVA showed condition had no significant effect on feelings of a-motivation ($F(2, 35) = 0.83, p = .44$): Interjections ($M = 3.07, SD = 1.19$); Music ($M = 3.72, SD = 1.67$); Control ($M = 3.55, SD = 1.10$).

4.3 Cognitive Workload

Two participants did not complete the Workload Profile questionnaire correctly and so their data was not included in the analysis for cognitive workload. In general, one-way ANOVA indicated that condition had no effect on total workload (summation of individual dimensions to provide an overall workload rating; $F(2, 33) = 0.44, p = .65$). Independent Kruskal-Wallis tests were also conducted to examine the differences in each individual workload dimension separately. No dimension showed a significant difference between conditions. As overall workload was not affected by interjections or music, and neither were the *Verbal*: "How much attention was required for verbal material (e.g., reading, processing linguistic material, listening to verbal conversations)?", and *Auditory*: "How much attention was required for executing the task based on the information auditorily received (ears)?" dimensions, considered most likely to be impacted, these findings provide support that expressive auditory gestures in pedagogical agents can be implemented without being detrimental to cognitive resource availability.

4.4 Teaching Behaviour

Lastly, participants' responses to the agent's on-topic questions (i.e., teaching statements) were analyzed. The responses were categorized into three answer types:

- *non-informational* answers – answers providing no new information or knowledge (these included, acknowledgement: e.g., "yes", "uh-huh"; agreement: e.g., "that's it"; maybe/ unknown: e.g., "something like that", "I don't know"; and rejection/ disagreement: e.g., "that's not it", "no"),
- *informational word-for-word* answers – answers containing information read word-for-word from the text provided, and
- *informational rephrase/reformulate* answers – answers providing information that was rephrased from the text or reformulated into a question.

To compare the proportion of each answer type (calculated as the number of times an answer type was given divided by the total number of answers given), one-way ANOVAs and linear models with step-wise model selection were used. Condition was not found to have an effect on the proportion of giving non-informational, informational word-for-word, or information rephrase/reformulate answers.

Non-informational. Through step-wise linear regression, we did however find that participants who reported being more interested in conversational agents gave a smaller proportion of non-informational answers in the Interjections condition than in the Control ($\beta = -0.12, t(26) = -2.93, p = .007$) and Music ($\beta = -0.15, t(26) = -3.47, p = .002$) conditions, whereas participants with more prior knowledge of the topic being taught gave a higher proportion of non-informational answers in the Interjections condition compared to the Control ($\beta = 0.12, t(26) = 2.62, p = .01$) and Music ($\beta = 0.08, t(26) = 2.21, p = .04$) conditions.

Informational word-for-word. Another linear model indicated that women gave a higher proportion of word-for-word answers in the Interjections condition compared to Music ($\beta = 0.14, t(32) = 2.07, p = .05$).

Informational rephrase/reformulate. Lastly, through step-wise linear regression, the resulting model indicated that participants that reported more interest in conversational agents gave a higher proportion of rephrase/reformulate answers in the Interjections condition than in the other two conditions: Music ($\beta = 0.24, t(29) = 4.69, p < .001$), Control ($\beta = 0.11, t(29) = 2.26, p = .03$). Conversely, these participants gave a lower proportion of rephrase/reformulate answers in the Music condition than in the Control ($\beta = -0.13, t(29) = -2.49, p = .02$).

5 DISCUSSION

Prior work suggests the importance of the relationship between tutee and tutor for increasing learning gains for the tutor (e.g., [47]) and that the expression of emotional and/or cognitive states can strengthen the human-agent relationship (e.g., [1, 53]). The purpose of this study was to understand how learners – in the role of tutor – perceive a voice-based agent – in the role of tutee – that adds expressive auditory gestures (a lesser studied form of expression in conversational systems) to its synthetic speech, and what effects these expressions have on the interaction, as well as learning outcomes. Numerous terminologies have been used to group and distinguish the various sounds and gestures that aim to express emotion through the auditory channel, including 'affect bursts' [55], 'anthropomorphic auditory icons' [56], and 'semantic-free utterances' (such as, non-linguistic, paralinguistic, and musical expressions [67]), that describe types of non-verbal cues, and 'speechcons' [4] and 'emotionally expressive interjections' [13] as variations of verbal expressions. In this study we focused on two specific auditory gestures: interjections (a type of verbal expression) and music (a type of non-verbal expression).

To investigate the perceived relationship with the agent, we measured Quality of Interaction as well as Rapport (along two dimensions: Emotional and Understanding), as both are suggested to contribute positively to learning outcomes. Researchers have previously investigated influencing rapport through strategies such as giving responses with appropriate emotional coloring [1], using off-topic dialogue [21], and entrainment [35, 36]. Our results suggest that adding expressive auditory gestures, especially in the form of interjections, aids in conveying affective state to the user. In the context of the user taking on the role of teacher, this perception of affective state may result in users having a better understanding of the agent as a learner or eliciting a form of emotional contagion, thereby enhancing perceived Emotional Rapport with the agent. This is of particular importance with regards to learning, as building Emotional Rapport can lead to feelings of emotional support and perception of a positive learning environment [18]. It further supports prior research that suggest the effects of interjections as a 'socio-affective glue' between interaction participants (e.g., [53]), and indicates its usefulness in educational contexts as well.

The lack of influence of interjections and musical sounds on improving Understanding Rapport or ratings of interaction quality, may be a result of the short interaction time in the study, as rapport can take time to develop [59], or it may be a consequence of the way in which we chose to measure and define rapport (i.e., self-report vs. behavioural or objective measures; Paralinguistic Rapport [44] vs. Virtual Rapport [19] or Natural Rapport [59]). It may also be due to the agent being voice-based, as prior studies (e.g., [66]) have shown that mixing two modalities – sound and facial expressions – can lead to stronger emotion recognition than when presented individually. This was emphasized by two participants in our study who mentioned that we could give the agent *"a face to make her more personable"* (I14) and to *"try bringing [the agent] on screen in the form of a figure to visualize better"* (M11). However, the results from this study indicate that for both interjections and musical expressions, the auditory gesture alone can be enough to convey the intended positive or negative affect. These findings are promising, as they imply that interjections and music can be used effectively in voice-only systems, without the need for a visual representation. The results also suggest that the MEB dataset [48], although not originally designed for the purpose in which it was used in this study (i.e., affective and cognitive expression of a voice-based conversational agent), can be used to express positive and negative valence in this scenario and be perceived as intended. This is of importance as much research is being undertaken to explore how expressive auditory gestures should and can be designed for conversational agents, often considered constrained to a certain context. Our findings indicate that musical expressions may be generalizable and could be a valuable direction for future work.

In terms of learning outcomes, we focused on the affective and cognitive outcomes, motivation and recall. Interjections were found to positively influence overall intrinsic and extrinsic motivation in the task, while brief expressive musical executions, although not resulting in increased feelings of rapport compared to the Control, lead to increases in certain dimensions of intrinsic and extrinsic motivation. Both types of motivation are considered important for successful learning [51], and therefore these findings are particularly promising for the use of such expressive auditory gestures in pedagogical agents. With regards to the cognitive learning outcome, though we found significantly higher reported levels of Emotional Rapport and motivation in participants in the Interjections condition compared to the Control, we did not find improvements in recall. This may be due to participants hitting a ceiling on learning gains, as pre-quiz scores, on average, were relatively high. However, as research proposes that rapport between human and agent leads to learning gains, our findings suggest a separation between rapport dimensions whereby developing Understanding Rapport is necessary for promoting cognitive learning outcomes, while feelings of Emotional Rapport enhance affective learning outcomes, such as motivation.

Although the human-like behaviour of adding expressive auditory gestures had positive impacts, they also lead to negative emotional and motivational responses in some learners, as they were perceived as *"creepy"* (I01) or *"scary"* (M05). Believability is an important feature of conversational agents and therefore diminished believability can influence the learner's experience and teaching behaviour overall. Beyond increasing the human-likeness of (teachable) pedagogical agents – a common goal of research in this area – future work could investigate at what point the human-likeness interferes with learning outcomes, and how exactly such types of auditory expressions are perceived along numerous dimensions including, for example whether they are perceived as spontaneous or as conveying intent (similar to children's perceptions of various auditory expressions [68]), and how this influences believability and in-turn learning outcomes. On the other hand, music circumvents the human-like sound of interjections, is not tied to any specific voice, and offers separate benefits to learning. Moreover, similar to interjections, our results indicate that music can aid in improving perception of intended affect, though it was not as effective as interjections at enhancing the perceived relationship and overall motivation. This indicates an opportunity for further exploration in this space. However, as suggested by prior work (e.g., [5]), as expressive sounds become more dissimilar to the voice, it can also result in the expressive auditory gesture and the voice being perceived as disjointed. Although this was not a common finding in this study, it should be taken into consideration in future work.

The influence of the user's personality, such as extraversion, and characteristics, such as gender, on the human-agent relationship is becoming increasingly elucidated by research in both learning and non-learning contexts [9, 10, 36, 58]. In fact, beyond condition effects, our results similarly demonstrate how certain characteristics can influence various of our measures. With regards to motivation, prior knowledge of the topic in the task had a negative influence on intrinsic motivation associated with wanting to learn, explore or understand something new, as well as a negative influence on extrinsic motivation related to personal growth, when participants interacted with the agents using either expressive auditory gesture. This is of particular importance in learning-by-teaching scenarios, whereby the user is to be the knowledge provider, i.e., teacher, but also learn in the process. It suggests that interacting with such an agent may be more effective for learners that themselves are not yet well-versed in the topic, as the expressive auditory gestures help motivate these learners, possibly by directing their learning or maintaining their engagement with the task. Our results also indicate that the amount of interest towards conversational agents that users expressed having prior to participating in the study can have a significant positive impact on their teaching behaviour during the interaction with the Interjections agent, while it had a negative influence when the agent used Music – possibly because it violates users' expectations of how voice-based agents should sound. As the teaching behaviour is a crucial aspect of the learning-by-teaching phenomenon, these findings are relevant to the design of such systems. However, as the number of participants in this study is relatively small, the results are preliminary and future work can investigate this more deliberately, along with taking personality characteristics into consideration as well, to further our understanding of the effects.

Overall, the results of this study support the viewpoint that, for pedagogical teachable agents to be successful, they need more than sophisticated technical capabilities; by displaying expressive behaviours they can have positive emotional, cognitive, and/or behavioural impacts.

6 FUTURE WORK AND LIMITATIONS

The role of the user in the conversation, as well as the conversational context, are important considerations when evaluating the use of expressive auditory gestures. In contrast to prior work, and the common role of voice-based agents i.e., as assistants, the role of the user in our study was an informer/teacher/tutor. This provides the context for the results, and there is potential for future work to investigate whether the results generalize to other roles, contexts, and agents as well. Furthermore, future work can examine how valence (positive vs. negative) can further affect learning, e.g., only positive when learning about one topic and only negative when learning about another. Additionally, as others have done in more socially-oriented conversations, e.g., Cohn et al. [13], future studies can look to understand what effects filler words, or a combination of fillers and interjections, or even interjections and music, can have on the constructs measured in this study. The results should also be interpreted in light of the Wizard of Oz technique used in the study. To mitigate the methodological and engineering concerns regarding the use of this technique, the Wizard was designed in a rigorous and replicable manner so as to facilitate the transition to more autonomous and complex systems in the future. We also acknowledge the small sample size and limited interaction time with the agent, making it difficult to generalize the results to larger or different populations. To handle small sample size, we avoided including more covariates in the ANOVA and selected the simplest possible model to explain our data in order to prevent overfitting.

Lastly, we recognize that the results obtained may be limited to the context and type of voice used (child-like). However, the findings

contribute to our understanding of the interaction and relationship building between learners and agents, suggesting how expressive auditory gestures can be used to influence learners' behavioural and affective experiences.

7 CONCLUSION

In this work, we present a study in which expressive auditory interjections and musical sounds were added to the dialogue of a teachable voice-based conversational agent. Where previous studies have focused largely on expressions through facial, gestural, and utterance-level cues, our work explores the lesser researched modality of auditory gestures – and investigates both language-dependent (interjection) and language-independent (music) expressions of cognitive and affective state. Measures of rapport and interaction quality were used to gauge the learner-agent relationship, with interjections leading to a significantly stronger sense of emotional connection with the agent, accompanied by increases in both intrinsic and extrinsic motivation towards putting in effort in the task, compared to the control agent with no added expressive auditory gesture. Musical executions on the other hand, were not found to lead to any significant increases in reported interaction quality or rapport compared to the control, but did result in increases in certain dimensions of motivation. Additionally, the design of pedagogical agents requires consideration of cognitive resource availability, which in-turn can influence learning outcomes, and the results suggest that both expressive auditory gesture can be implemented without imposing extraneous cognitive workload.

As interjections are highly prevalent in human-human dialogue and are becoming more widely available in popular systems, e.g., *"speechcons"* in the Amazon Alexa, our results provide evidence for the influence of interjections on human-agent relationship building, highlighting differences between emotional and understanding rapport, as well as presenting effects of interjections on motivation. The findings highlight practical insights for voice-system designers in education as well as across other domains including healthcare, entertainment, and customer service, for example, in which building rapport and enhancing motivation can be of similar value.

ACKNOWLEDGMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

REFERENCES

[1] Jaime C. Acosta and Nigel G. Ward. 2011. Achieving Rapport with Turn-by-Turn, User-Responsive Emotional Coloring. *Speech Commun.* 53, 9–10 (nov 2011), 1137–1148. https://doi.org/10.1016/j.specom.2010.11.006

[2] Alan Agresti. 2010. *Analysis of ordinal categorical data: Second Edition.* John Wiley & Sons.

[3] James L. Alty, Dimitrios Rigas, and Paul Vickers. 2005. Music and speech in auditory interfaces: When is one mode more appropriate than another?. In *International Conference on Auditory Display*. 351–357.

[4] Amazon. 2021. *Speechcons (Interjections).* https://developer.amazon.com/en-US/docs/alexa/custom-skills/speechcon-reference-interjections.html

[5] Matthew P. Aylett, Yolanda Vazquez-Alvarez, and Skaiste Butkute. 2020. Creating Robot Personality: Effects of Mixing Speech and Semantic Free Utterances. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 110–112. https://doi.org/10.1145/3371382.3378330

[6] Amy L. Baylor. 2011. The design of motivational agents and avatars. *Educational Technology Research and Development* 59, 2 (2011), 291–300. https://doi.org/10.1007/s11423-011-9196-3

[7] Amy L. Baylor and Soyoung Kim. 2009. Designing nonverbal communication for pedagogical agents: When less is more. *Computers in Human Behavior* 25, 2 (2009), 450–457. https://doi.org/10.1016/j.chb.2008.10.008

[8] Pascal Belin, Robert J. Zatorre, and Pierre Ahad. 2002. Human temporal-lobe response to vocal sounds. *Cognitive Brain Research* 13, 1 (2002), 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2

[9] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. 2014. How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits. In *Park H.S., Salah A.A., Lee Y.J., Morency LP., Sheikh Y., Cucchiara R. (eds) Human Behavior Understanding. HBU 2014*, Vol. 8749, Lecture Notes in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-11839-0_1

[10] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. 2017. Rapport with Virtual Agents: What Do Human Social Cues and Personality Explain? *IEEE Transactions on Affective Computing* 8, 3 (2017), 382–395. https://doi.org/10.1109/TAFFC.2016.2545650

[11] Catherine C. Chase, Doris B. Chin, Marily A. Oppezzo, and Daniel L. Schwartz. 2009. Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology* 18, 4 (2009), 334–352. https://doi.org/10.1007/s10956-009-9180-4

[12] Richard E. Clark and Sunhee Choi. 2005. Five design principles for experiments on the effects of animated pedagogical agents. *Journal of Educational Computing Research* 32, 3 (2005), 209–225. https://doi.org/10.2190/7LRM-3BR2-44GW-9QQY

[13] Michelle Cohn, Chun-Yen Chen, and Zhou Yu. 2019. A Large-Scale User Study of an Alexa Prize Chatbot: Effect of TTS Dynamism on Perceived Quality of Social Dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Stockholm, Sweden, 293–306. https://doi.org/10.18653/v1/W19-5935

[14] Edward L. Deci and Richard M. Ryan. 1985. *Intrinsic motivation and self-determination in human behavior.* Springer Science & Business Media.

[15] Pieter M. A. Desmet, Martijn H. Vastenburg, and Natalia Romero. 2016. Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research* 14, 3 (2016), 241–279.

[16] David Duran. 2017. Learning-by-teaching. Evidence and implications as a pedagogical mechanism. *Innovations in Education and Teaching International* 54, 5 (2017), 476–484. https://doi.org/10.1080/14703297.2016.1156011

[17] Laura Ferreri and Laura Verga. 2016. Benefits of music on verbal learning and memory: How and when does it work? *Music Perception: An Interdisciplinary Journal* 34, 2 (2016), 167–182. https://www.jstor.org/stable/10.2307/26417442

[18] Brandi N. Frisby. 2019. The influence of emotional contagion on student perceptions of instructor rapport, emotional support, emotion work, valence, and cognitive learning. *Communication Studies* 70, 4 (2019), 492–506. https://doi.org/10.1080/10510974.2019.1622584

[19] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating Rapport with Virtual Agents. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '07)*. Springer-Verlag, Berlin, Heidelberg, 125–138. https://doi.org/10.1007/978-3-540-74997-4_12

[20] Ivan Gris Sepulveda. 2015. *Physical engagement as a way to increase emotional rapport in interactions with embodied conversational agents.* Ph.D. Dissertation. The University of Texas at El Paso.

[21] Agneta Gulz, Magnus Haake, and Annika Silvervarg. 2011. Extending a Teachable Agent with a Social Conversation Module: Effects on Student Experiences and Learning. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (Auckland, New Zealand) *(AIED'11)*. Springer-Verlag, Berlin, Heidelberg, 106–114.

[22] Agneta Gulz, Annika Silvervarg, and Björn Sjödén. 2010. Design for Off-Task Interaction - Rethinking Pedagogy in Technology Enhanced Learning. In *Proceedings of the 2010 10th IEEE International Conference on Advanced Learning Technologies (ICALT '10)*. IEEE Computer Society, USA, 204–206. https://doi.org/10.1109/ICALT.2010.63

[23] Jiaxiong Hu, Yun Huang, Xiaozhu Hu, and Yingqing Xu. 2021. *Enhancing the Perceived Emotional Intelligence of Conversational Agents through Acoustic Cues.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451660

[24] Eun-Sook Jee, Yong-Jeon Jeong, Chong Hui Kim, and Hisato Kobayashi. 2010. Sound design for emotion and intention expression of socially interactive robots. *Intelligent Service Robotics* 3 (2010), 199–206. https://doi.org/10.1007/s11370-010-0070-7

[25] Eun-Sook Jee, Chong Hui Kim, Soon-Young Park, and Kyung-Won Lee. 2007. Composition of Musical Sound Expressing an Emotion of Robot Based on Musical Factors. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*. 637–641. https://doi.org/10.1109/ROMAN.2007.4415161

[26] Eun-Sook Jee, Soon-Young Park, Chong Hui Kim, and Hisato Kobayashi. 2009. Composition of musical sound to express robot's emotion with intensity and

synchronized expression with robot's behavior. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. 369–374. https://doi.org/10.1109/ROMAN.2009.5326258

[27] 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.

[28] Patrik N. Juslin and Petri Laukka. 2004. Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *Journal of New Music Research* 33, 3 (2004), 217–238. https://doi.org/:10.1080/0929821042000317813

[29] Kurt Kraiger, J. Kevin Ford, and Eduardo Salas. 1993. Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology* 78, 2 (1993), 311–328. https://doi.org/10.1037//0021-9010.78.2.311

[30] David R. Krathwohl, Benjamin S. Bloom, and Bertram B Masia. 1964. *Taxonomy of educational objectives, Handbook II: Affective domain.* David McKay Company, Inc.

[31] Jong-Eun Roselyn Lee, Clifford Nass, Scott Brenner Brave, Yasunori Morishima, Hiroshi Nakajima, and Ryota Yamada. 2007. The case for caring colearners: The effects of a computer-mediated colearner agent on trust and learning. *Journal of Communication* 57 (2007), 183–204. https://doi.org/10.1111/j.1460-2466.2007.00339.x

[32] Krittaya Leelawong and Gautam Biswas. 2008. Designing Learning by Teaching Agents: The Betty's Brain System. *International Journal of Artificial Intelligence in Education* 18, 3 (aug 2008), 181–208.

[33] James Lester, Karl Branting, and Bradford Mott. 2004. Conversational agents. In *The Practical Handbook of Internet Computing*. Chapman & Hall CRC, London.

[34] Tze Wei Liew, Nor Azan Mat Zin, and Noraidah Sahari. 2017. Exploring the Affective, Motivational and Cognitive Effects of Pedagogical Agent Enthusiasm in a Multimedia Learning Environment. *Human-Centric Computing and Information Sciences* 7, 1, Article 89 (dec 2017), 21 pages. https://doi.org/10.1186/s13673-017-0089-2

[35] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2016. Effects of Voice-Adaptation and Social Dialogue on Perceptions of a Robotic Learning Companion. In *The 11th ACM/IEEE International Conference on Human Robot Interaction* (Christchurch, New Zealand) *(HRI '16)*. IEEE Press, 255–262. https://doi.org/10.1109/HRI.2016.7451760

[36] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2021. Effects of adapting to user pitch on rapport perception, behavior, and state with a social robotic learning companion. *User Modeling and User-Adapted Interaction* 31, 1 (2021), 35–73. https://doi.org/10.1007/s11257-020-09267-3

[37] Nicola Mammarella, Beth Fairfield, and Cesare Cornoldi. 2007. Does music enhance cognitive performance in healthy older adults? The Vivaldi effect. *Aging Clinical and Experimental Research* 19, 5 (2007), 394–399. https://doi.org/10.1007/BF03324720

[38] Dominic W. Massaro. 2006. Embodied agents in language learning for children with language challenges. In *Miesenberger K., Klaus J., Zagler W.L., Karshmer A.I. (eds) Computers Helping People with Special Needs. ICCHP 2006.*, Vol. 4061, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11788713_118

[39] Dominic W. Massaro, Ying Liu, Trevor H. Chen, and Charles Perfetti. 2006. A multilingual embodied conversational agent for tutoring speech and language learning. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP, September, Pittsburgh, PA)*. Universität Bonn, Bonn, Germany, 825–828.

[40] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an Intelligent Agent through Wizard of Oz. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 277–284. https://doi.org/10.1145/169059.169215

[41] Peter McCullagh and J.A. Nelder. 1989. *Generalized Linear Models. 2nd Ed.* Chapman & Hall, London.

[42] Michael Neff, Yingying Wang, Rob Abbott, and Marilyn Walker. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *Allbeck J., Badler N., Bickmore T., Pelachaud C., Safonova A. (eds) Intelligent Virtual Agents. IVA 2010. Lecture Notes in Computer Science*, Vol. 6356. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15892-6_24

[43] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Haizhou Li, and Swee Lan See. 2013. Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy. *International Journal of Social Robotics* 5, 2 (2013), 171–191. https://doi.org/10.1007/s12369-012-0171-x

[44] David Novick and Iván Gris. 2014. Building rapport between human and ECA: A pilot study. In *Kurosu M. (eds) Human-Computer Interaction. Advanced Interaction Modalities and Techniques. HCI 2014. Lecture Notes in Computer Science*, Vol. 8511. Springer, Cham. https://doi.org/10.1007/978-3-319-07230-2_45

[45] Benjamin D. Nye, Arthur C. Graesser, and Xiangen Hu. 2014. AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education* 24 (2014), 427–469. https://doi.org/10.1007/s40593-014-0029-5

[46] Amy Ogan, Samantha Finkelstein, Elijah Mayfield, Claudia D'Adamo, Noboru Matsuda, and Justine Cassell. 2012. "Oh Dear Stacy!": Social Interaction, Elaboration, and Learning with Teachable Agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/2207676.2207684

[47] Amy Ogan, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012. Rudeness and Rapport: Insults and Learning Gains in Peer Tutoring. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (Chania, Crete, Greece) *(ITS'12)*. Springer-Verlag, Berlin, Heidelberg, 11–21. https://doi.org/10.1007/978-3-642-30950-2_2

[48] Sébastien Paquette, Isabelle Peretz, and Pascal Belin. 2013. The "Musical Emotional Bursts": a validated set of musical affect bursts to investigate auditory affective processing. *Frontiers in Psychology* 4, 509 (2013). https://doi.org/10.3389/fpsyg.2013.00509

[49] Aditi Ramachandran, Chien-Ming Huang, Edward Gartland, and Brian Scassellati. 2018. Thinking Aloud with a Tutoring Robot to Enhance Learning. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3171221.3171250

[50] Rod D. Roscoe and Michelene T.H. Chi. 2007. Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research* 77, 4 (2007), 534–574.

[51] Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology* 25, 1 (2000), 54–67. https://doi.org/10.1006/ceps.1999.1020

[52] Martin Saerbeck, Tom Schut, Christoph Bartneck, and Maddy D. Janse. 2010. Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1613–1622. https://doi.org/10.1145/1753326.1753567

[53] Yuko Sasa and Véronique Auberge. 2014. Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the "socio-affective glue". *Speech Prosody* 7 (2014). https://doi.org/10.21437/SpeechProsody.2014-5

[54] E. Glenn Schellenberg and Michael W. Weiss. 2013. Music and cognitive abilities. In *The psychology of music*, Diana Deutsch (Ed.). Elsevier Academic Press, 499–550. https://doi.org/10.1016/B978-0-12-381460-9.00012-2

[55] Klaus R. Scherer. 1994. Affect Bursts. In *Emotions: Essays on Emotion Theory*, Joseph A. Sergeant Stephanie H. M. van Goozen, Nanne E. van de Poll (Ed.). Psychology Press, 175–208.

[56] Michael Schmitz, Benedict C.O.F. Fehringer, and Mert Akbal. 2015. Expressing Emotions With Synthetic Affect Bursts. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (London, United Kingdom) *(CHI PLAY '15)*. Association for Computing Machinery, New York, NY, USA, 91–95. https://doi.org/10.1145/2793107.2793139

[57] Dale H. Schunk. 1991. Self-efficacy and academic motivation. *Educational Psychologist* 26, 3-4 (1991), 207–231.

[58] Betty Tärning, Agneta Gulz, Magnus Haake, et al. 2019. Instructing a teachable agent with low or high self-efficacy–does similarity attract? *International Journal of Artificial Intelligence in Education* 29, 1 (2019), 89–121. https://doi.org/10.1007/s40593-018-0167-2

[59] Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological Inquiry* 1, 4 (1990), 285–293. https://doi.org/10.1207/s15327965pli0104_1

[60] Pamela S. Tsang and Velma L. Velazquez. 1996. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics* 39, 3 (1996), 358–381. https://doi.org/10.1080/00140139608964470

[61] Robert J. Vallerand, Luc G. Pelletier, Marc R. Blais, Nathalie M. Briere, Caroline Senecal, and Evelyne F. Vallieres. 1992. The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement* 52, 4 (1992), 1003–1017. https://doi.org/10.1177/0013164492052004025

[62] Michelle M.E. Van Pinxteren, Mark Pluymaekers, and Jos G.A.M. Lemmink. 2020. Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management* 31, 2 (2020), 203–225. https://doi.org/10.1108/JOSM-06-2019-0175

[63] Paul Vickers and James L. Alty. 2002. Using music to communicate computing information. *Interacting with Computers* 14, 5 (2002), 435–456. https://doi.org/10.1016/S0953-5438(02)00003-6

[64] Symeon P. Vlachopoulos and Costas I. Karageorghis. 2005. Interaction of external, introjected, and identified regulation with intrinsic motivation in exercise: relationships with exercise enjoyment. *Journal of Applied Biobehavioral Research* 10, 2 (2005), 113–132. https://doi.org/10.1111/j.1751-9861.2005.tb00007.x

[65] Wayne Ward, Ronald Cole, Daniel Bolaños, Cindy Buchenroth-Martin, Edward Svirsky, Sarel Van Vuuren, Timothy Weston, Jing Zheng, and Lee Becker. 2011.

My Science Tutor: A Conversational Multimedia Virtual Tutor for Elementary School Science. *ACM Transactions on Speech and Language Processing* 7, 4 (2011), 1–29. https://doi.org/10.1145/1998384.1998392

[66] Selma Yilmazyildiz, David Henderickx, Bram Vanderborght, Werner Verhelst, Eric Soetens, and Dirk Lefeber. 2013. Multi-modal emotion expression for affective human-robot interaction. In *Proceedings of the Workshop on Affective Social Speech Signals (WASSS 2013), Grenoble, France.*

[67] Selma Yilmazyildiz, Robin Read, Tony Belpeame, and Werner Verhelst. 2016. Review of semantic-free utterances in social human–robot interaction. *International Journal of Human-Computer Interaction* 32, 1 (2016), 63–85. https://doi.org/10.1080/10447318.2015.1093856

[68] Jason C. Yip, Kiley Sobel, Xin Gao, Allison Marie Hishikawa, Alexis Lim, Laura Meng, Romaine Flor Ofiana, Justin Park, and Alexis Hiniker. 2019. Laughing is Scary, but Farting is Cute: A Conceptual Model of Children's Perspectives of Creepy Technologies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300303