NATURALNESS AND RAPPORT IN A PITCH ADAPTIVE LEARNING COMPANION

Nichola Lubold [1], *Heather Pon-Barry* [2], *Erin Walker* [1]

[1] CIDSE, Arizona State University, Tempe, AZ, USA
[2] Department of Computer Science, Mount Holyoke College, South Hadley, MA, USA

ABSTRACT

Observed frequently in human-human interactions, entrainment is a social phenomenon in which speakers become more like each other over the course of a conversation. Acoustic-prosodic entrainment occurs when individuals adapt their acoustic-prosodic speech features, such as pitch and intensity. Correlated with communicative success, naturalness, and conversational flow as well as social variables such as rapport, a dialogue system which automatically entrains has the potential to improve verbal interactions by increasing rapport, naturalness, and conversational flow. In an application like the learning companion, such a socially responsive dialogue system may improve learning and motivation. However, it is not clear how to produce entrainment in an automatic dialogue system in ways that produce the effects seen in human-human dialogue. In this paper, we take the first steps towards implementing a spoken dialogue system which can entrain. We propose three methods of pitch adaptation based on analysis of human entrainment, and design and implement a system which can manipulate the pitch of text-to-speech output adaptively. We find a clear relationship between perceptions of rapport and different forms of pitch adaptations. Certain adaptations are perceived as significantly more natural and rapport-like. Ultimately, adapting by shifting the pitch contour of the text-to-speech output by the mean pitch of the user results in the highest reported measures of rapport and naturalness.

***Index Terms*—** pitch, adaptation, dialogue system, naturalness, rapport

1. INTRODUCTION

Spoken dialogue systems are a part of mainstream society, from automated answering systems to the advent of voiced personal assistants such as Siri, Google Now, and more recently, Cortana. With advances made in automatic speech recognition (ASR), we've seen the ability of these technologies to hold a conversation improve considerably. As dialogue systems become more pervasive, there is an increasingly important role for socially responsive dialogue systems that can effectively socially engage the user.

One application where a socially responsive dialogue system is potentially very impactful is the learning companion [4]. A learning companion provides a support system for students with the goal of improving learning by providing both task-related feedback and motivational support. Learning companions, based on the theory that learning is influenced by social interactions [32], require social sensitivity to influence students' socio-motivational factors and increase student learning. In the case of learning companions, the ability to be socially responsive positively impacts learning (e.g., [14]).

In this paper, we are interested in utilizing acoustic-prosodic features of speech to improve the social responsiveness of a learning companion's dialogue system. We explore the phenomenon acoustic-prosodic entrainment as a possible mechanism. Acoustic-prosodic entrainment is where two speakers adapt their acoustic-prosodic features including their tone, intensity, and speaking rate to mirror one another [17]. Correlated with a number of factors including communicative success, conversational flow, and social factors like rapport, acoustic-prosodic entrainment is critical to the naturalness and flow of dialogue [1, 19, 23]. A dialogue system which automatically entrains has the potential to improve verbal interactions by increasing the above factors. In a learning companion, such a dialogue system may improve learning and motivation as well.

To this end, we design and implement a system which adapts the pitch of a text-to-speech (TTS) output real-time, to accommodate to the pitch of the user. We implement three different forms of pitch adaptation, inspired by how human conversational partners entrain. We collect data from four individuals interacting with each form of pitch adaptation, and then, using crowd-sourced analysis via Amazon Mechanical Turk, we compare the different adaptations to the baseline TTS on rapport and naturalness perceived by third-party observers. We find the most effective adaptation is on pitch mean where the standard TTS pitch contour is maintained but the contour is adapted to the average pitch of the user. These findings provide insight into how pitch adaptations are perceived overall and how to proceed with creating effective entrainment in the dialogue system of a learning companion.

In the below sections, we motivate this work in relation to entrainment, rapport, and learning, and describe related work on manipulating acoustic prosodic features. In Section 2, we describe in more detail the pitch adaptations performed and our hypotheses. Section 3 outlines the intended application and the dialogue interface we built to implement the pitch adaptations. Section 4 outlines the process and results of evaluating the naturalness and rapport of these adaptations. We discuss the implications of these results in Section 5, concluding with a brief discussion and thoughts on future work in Section 6.

1.1 Entrainment, Rapport, & Learning

Entrainment, known also as accommodation, occurs when dialogue partners adapt their behavior to each other during an interaction. Entrainment can be gestural, via gaze or facial expressions [15], word-based or lexical [8], or speech-based [26]. In this paper, we are interested specifically in acoustic-prosodic entrainment, when two speakers adapt their acoustic-prosodic speech features to one another, and we focus on one acoustic-prosodic feature – pitch.

Entrainment on pitch prominently differentiates communicative success [1] and entrainment measures derived from pitch features are significantly higher in positive interactions [16]. In addition, entrainment on pitch is linked to both rapport and learning [30, 20]. In theories of learning, it has been proposed that all learning is social [32] and feelings of rapport have been shown to impact how much students learn from interactions with learning companions [25]. By entraining on pitch, a learning companion may increase rapport, and the increased rapport may improve learning gains. While other forms of automated entrainment may also benefit learners, our focus here is on assessing an optimal strategy for pitch adaptation.

1.2 Manipulating Acoustic-Prosodic Features

Manipulating the acoustic-prosodic features of the text-to-speech output of a dialogue system to influence entrainment has precedence in past work. These efforts were focused on features which are easy to manipulate, such as intensity and speaking rate [28, 29, 5]. In these scenarios, the acoustic-prosodic features were adjusted in order to transform the overall dialogue output without regard to the human speaker. The results show that humans will entrain to a computer, adapting their own voice to the computer. The manipulations did not explore the effect of computer adaptation.

Adapting the acoustic-prosodic features of the output of a spoken dialogue system to a user is a more recent innovation. In her thesis on entrainment in human-human and human-computer dialogue [18], Rivka Levitan appears to be the first to look at adapting text-to-speech on a turn-by-turn basis based on the user's acoustic-prosodic features. Levitan found that individuals interacting with a virtual agent which entrained on intensity and speaking rate unconsciously trusted that agent more than an agent which did not entrain on these features. This provides support that entrainment triggers social responses in line with traditional human-computer interaction theory, which suggests that humans respond socially to computers in similar ways as they respond to other humans [22].

For this work, we take a similar approach to Levitan in adapting acoustic-prosodic features on a turn-by-turn basis but our focus is on pitch. While pitch has been looked at for improving the naturalness of text-to-speech [6, 31, 33], it has received less attention as feature for automated entrainment. Given the history of entrainment on pitch, we hope to confirm that humans respond to an entraining computer in the same way they respond to entraining humans by finding that social variables correlated with human-human entrainment on pitch can be affected by computer adaptation. As the best way to entrain on pitch is not yet clear, we evaluate the success of different pitch adaptations in producing meaningful entrainment.

2. PITCH ADAPTATION METHODOLOGY

Three forms of pitch adaptation are proposed, inspired by observations of how human conversation partners entrain. Prosodic entrainment is often measured along multiple dimensions. We focus on *proximity*. Proximity measures how near the acoustic-prosodic features of two speakers are, on a turn-by-turn basis. It is the most frequent form of entrainment observed in turn-by-turn analyses, compared to other acoustic features and types of entrainment. Proximal entrainment on pitch specifically has been linked to greater rapport, communicative success, and positivity [19, 1, 15] so the adaptations we explore here focus on pitch.

The three proposed methods of pitch adaptation operate at the turn-level. The system adapts its pitch based on the estimated pitch values from the previous speaker's turn, as opposed to the longer dialogue history. Figure 1 illustrates the pitch contour of a sample waveform alongside three adaptations.

The first method of pitch adaptation is **mirror partner**. With mirror partner, we adapt the text to speech output to the entire pitch contour of the speaker's previous turn by replacing the original contour of the TTS with the contour of the speaker. To account and control for differences in utterance length, we resize the speaker's utterance to be the same length as the proposed text to speech output prior to applying the speaker's contour to the output. This approach to adaptation would maximize the level of entrainment, following the metrics used in past work [20].

Shift+contour is an alternative method of pitch adaptation that maintains the contour of the original TTS but shifts it up

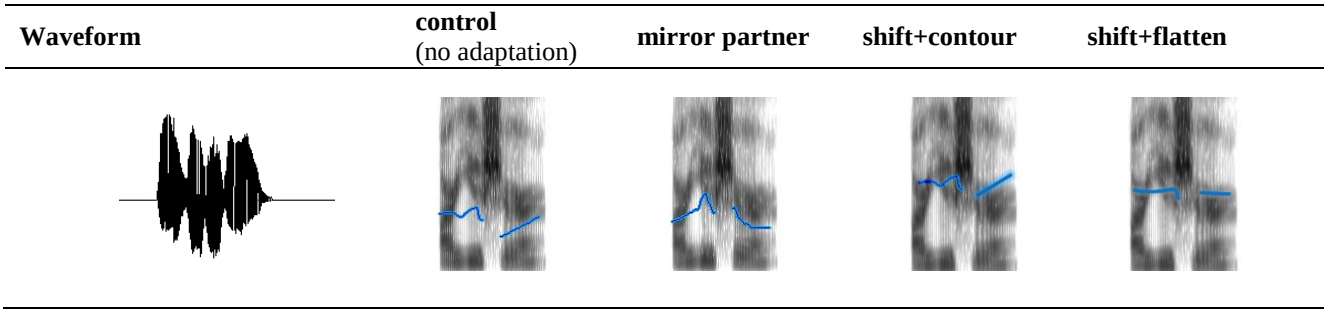

Fig. 1: Spectrograms and pitch contours of the synthesized waveforms (original + three with pitch adaptation).

or down to match the mean pitch of the speaker**.** While mirroring the shape of a partner's pitch contour might strengthen automated measures of entrainment, there is the possibility of "over-adaptation" and of a mismatch between pitch contour and syntactic and semantic structure. Since entrainment on pitch mean has been found to be correlated with learning and rapport, above and beyond other attributes of pitch, shift+contour only adapts the pitch mean.

We introduce a third adaptation called **shift+flatten**. This adaptation serves as a minimum manipulation baseline in respect to the other two approaches. Still adapting on a single feature, pitch mean, we flatten the pitch contour of the TTS to the pitch mean of the user. The TTS output maps to the average pitch of the student. As this adaptation is intuitively the least realistic, we would not expect it to produce more rapport than the other two conditions. Thus, it serves as a baseline comparison for the more sophisticated pitch adaptations we propose, in addition to the c**ontrol**, the original synthesized waveform with no adaptation on pitch.

2.2 Hypotheses

We hypothesize the pitch adaptations will result in more rapport than the basic text-to-speech. Specifically, we hypothesize mirror partner will produce more rapport than shift+contour or the basic text-to-speech baseline. The third adaptation, shift+flatten, will generate the least rapport. While we are interested in how rapport differs for different pitch adaptations, we want to ensure that the adaptations are perceived to be as natural as the original synthesized waveform. We hypothesize that mirror partner and shift+contour will not be significantly different from the baseline text-to-speech. We also hypothesize that shift+flatten will be significantly less natural.

3. LEARNING COMPANION APPLICATION

We implement a virtual learning companion to analyze the effect of the pitch adaptations. Students interact using spoken language with a virtual entraining agent, referred to as Quinn, and a web application. Quinn is present throughout the interaction on a tablet device. Quinn's facial expressions are animated when speaking, and neutral otherwise. Underlying the web application is a collection of variable equation problems (i.e. "Solve 4x +3y = 80 for x"). The application presents each problem separately and includes steps to reach a solution. The problems are ordered in increasing order of difficulty. Quinn and an example of the web interface display for step one of a sample problem are found in Figure 2.

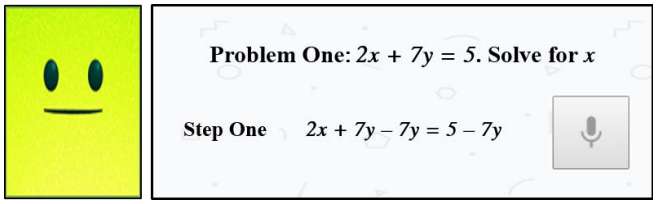

Fig. 2: Quinn and an example of the web interface display for step one of a sample problem.

Before teaching Quinn, students are given a sample problem to practice how to teach the problem. They are then introduced to Quinn and shown how to use the interface. To teach Quinn the problem, the student clicks on the microphone displayed on the page which enables real-time speech. They then proceed to walk Quinn verbally through the steps displayed on the screen. The speech interaction is real-time, and the dialogue is recorded as the student speaks. After explaining each step, the students are instructed to pause, giving Quinn a chance to respond. A sample of the dialogue taken from the present study is given below; '*Q*' represents Quinn and '*S*' represents the student.

S: We will divide both sides by negative six

Q: Can you explain why we divide?

S: On the left hand side, we have negative 6y. We need to have it equal just y so we need to get rid of the negative six. The easiest way is to divide.

Q: Thank you explaining! I get it now. So we divide. Then what?

3.1 Dialogue Interface

To explore the effect of pitch adaptations on perceptions of naturalness and rapport, we build a dialogue interface which can manipulate the pitch of the text-to-speech output. We

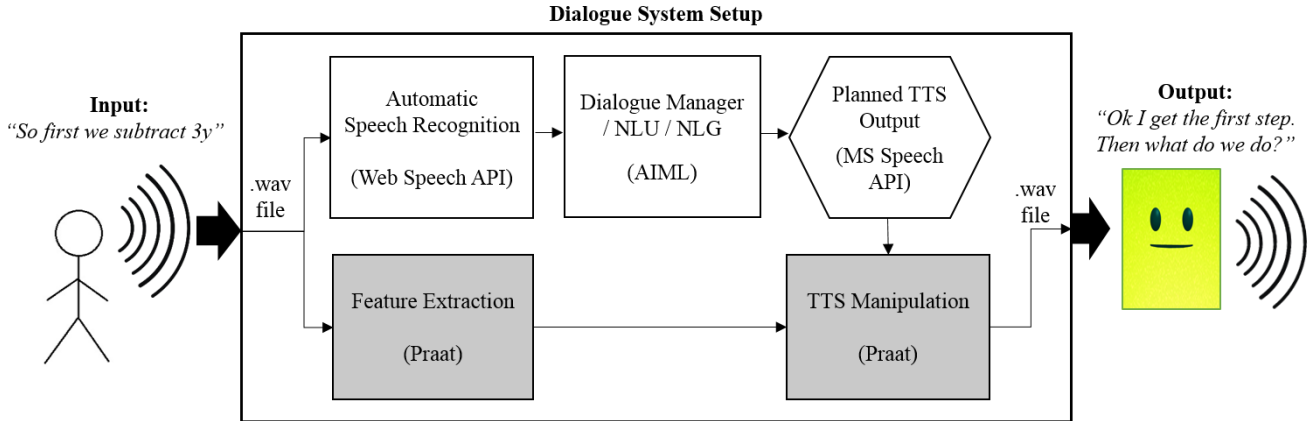

Fig. 3: Dialogue system for Quinn. The darker boxes indicate components belonging to the pitch adaptation module.

designed the system in a modular fashion so the pitch adaptation module can be introduced independently into other systems. Speech recognition was performed using the Web Speech API specification[1]. The dialogue manager was developed using Artificial Intelligence Markup Language (AIML) developed by Richard Wallace [34]. AIML is an XML-compliant pattern matching language. We utilized the tool PandoraBots[2] to develop the AIML. The text-to-speech output is produced with the Microsoft Speech API. The feature extraction and pitch adaptations[3] are implemented using Praat [2]. The basic dialogue system design and technologies utilized are illustrated in Figure 3. The darker boxes indicate components of the pitch adaptation module.

5. NATURALNESS & RAPPORT EVALUATION

5.1 Spoken Dialogue Data Collection

We collect 32 dialogues from four individuals. In each study, an undergraduate college student interacts with Quinn using the web application to teach Quinn how to solve eight variable equation math problems. For two problems, Quinn speaks with a non-transformed baseline speech. For the remaining six problems, Quinn alternates the type of adaptation for each problem. Two full problems are given for each type of adaptation; we collect each problem as a separate dialogue for a total of 8 dialogues per student. Statistics for the collected corpus are shown in Table 1. The gender of Quinn's voice was chosen to match the gender of the student. The four case studies are gender balanced with two males and two females. The gender of the speaker drove the gender of Quinn's voice. If the student was a female, then Quinn was female. If the student was male, Quinn was male.

Table 1: Dialogue and turn statistics for corpus

	Mean	Std. Dev.
Dialogue length (min)	5.4	2.1
Number of turns	30	10
Turn length (sec)	10.8	4.6

5.2 Amazon Mechanical Turk (AMT) Evaluation

Taking the dialogue corpus collected in section 5.1, we manually select 40 exchanges from each of the student-Quinn dialogues. An exchange is considered to be two adjacent turns by different speakers (i.e. the student and Quinn). We select ten exchanges for the baseline text-to-speech and ten exchanges for each of the three types of adaptation, focusing on those exchanges with maximum coherency and minimal pausing or silence, eliminating any exchanges were speech recognition may have failed. The ten exchanges are evenly split between two scenarios. In the first scenario, Quinn is the first speaker in the exchange. In the second scenario, Quinn is the second speaker and is responding. With a total of 40 exchanges per student, we utilized Amazon Mechanical Turk (AMT), a popular resource for crowdsourcing research tasks including annotations, transcripts, and subjective analysis [3]. We use AMT to obtain 10 random, perceptual evaluations per exchange for a total of 400 evaluations per student or 1600 evaluations. Using third party ratings such as those collected through AMT is a standard technique in the evaluation of naturalness and social features of dialogue systems [13]. In addition, avoiding first-person ratings allowed us to present all dialogue approaches to each of the four individuals without worrying about how their perceptions of one approach might affect their ratings of a different approach. Through AMT, individuals, referred to as workers, were asked to listen to each exchange and answer a series of questions[4] regarding the speakers. Each worker has access to

[1] https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html
[2] www.pandorabots.com
[3] Link to Praat script implementing the adaptations: http://www.public.asu.edu/~nlubold/Research/pitchadapt.praat
[4] http://www.public.asu.edu/~nlubold/Research/sampleHit.html

evaluate 160 exchanges (40 per student). To evaluate naturalness, we use Mean Opinion Score or MOS [12]. With MOS, workers are asked to evaluate the quality of the voice on a Likert scale of 1-5, where 1 is very poor and 5 is completely natural. Workers evaluated both the human speaker and Quinn on this scale.

For evaluating rapport, we adopt a subset of questions from the rapport scale utilized by [9] and [19]. Workers are asked the following two questions about the relationship between the speakers on a Likert scale of 1-5, where 1 is "not at all" and 5 is "a lot." In the questions below, Alex refers to the student and Quinn refers to the virtual agent. We selected these questions because they target a shared feeling between speakers. The responses are averaged for one rapport rating.

1) Alex and Quinn understood each other
2) There is a sense of closeness between Alex and Quinn

In total, 174 workers provided evaluations of the audio. 12% or 21 workers rated 30% or more of the possible 160 exchanges they had access to while 40% of the workers listened to and rated only one exchange. In analyzing the results below, we treat each rating as the unit of analysis.

5.3 Results

To analyze the effect of the pitch adaptations in terms of rapport and naturalness and evaluate our hypotheses, we run a basic statistical analysis of the relationship between type of adaptation, naturalness, and rapport. A subsequent in-depth analysis of the individual participants and differences in ratings reveals a connection between social content of exchanges, adaptation type, and degree of rapport perceived.

5.3.1. Naturalness & Rapport

Naturalness – We perform a one-way analysis of variance (ANOVA) with the type of adaptation (mirror partner, shift+contour, shift+flatten, and control) as a factor and naturalness as the dependent variable. Table 2 gives the means and standard deviations for each condition. The ANOVA analysis indicates statistically significant differences among type of adaptations, $F(3, 1599) = 19.9$, $p < 0.001$. Tukey post hoc tests indicate that shift+contour was perceived as significantly more natural than either mirror partner ($p < 0.001$) or shift-flatten ($p < 0.001$). We expected to find mirror partner and shift+contour were as natural as the control. We find mirror partner is perceived to be much less natural, on par with shift+flatten. We also find shift+contour is not significantly different from the control, where no adaptation is performed ($p = 0.52$). These results lead us to conclude that in pursing implementing an automatically entraining system, shift+contour, adapting pitch by shifting the TTS contour, is the most natural of the adaptations reviewed and is as natural as a non-manipulated TTS output.

Table 2: Means and standard deviations for naturalness on each pitch adaptation

	Mean	Std. Dev.
control	2.22	1.36
mirror partner	1.79	1.11
shift+contour	2.39	1.33
shift+flatten	1.85	1.15

Rapport – To identify differences in how rapport is perceived for each of the pitch adaptations, we perform a one-way ANOVA with adaptation type as a factor and rapport as the dependent variable. Table 3 gives the means and standard deviations. We find statistically significant differences among the types of adaptations, $F(3, 1599) = 5.63$, $p < 0.001$. Our hypothesis was that mirror partner would result in the most rapport, followed by shift+contour. Shift+flatten, we expected to be the lowest. Interestingly, we see shift+contour is on par with mirror partner. Both are indicating higher, equivalent degrees of rapport over the control. Using Tukey post hoc tests to analyze which of the pitch adaptations are significantly different, we find shift+contour generates significantly higher perceptions of rapport than shift+flatten ($p < 0.01$). Differences between shift+contour, mirror partner, and control are not significant.

Table 3: Means and standard deviations for rapport on each pitch adaptation

	Mean	Std. Dev.
control	3.56	1.17
mirror partner	3.73	1.07
shift+contour	3.74	1.07
shift+flatten	3.35	1.15

Given that rapport may have been influenced by several aspects of the human-agent dialogue beyond pitch adaptation, we pursue in the next section a more in depth analysis of the exchanges, to identify whether there are any further conclusions we can draw regarding the social factors introduced by the pitch adaptations.

5.3.2. Identifying Moderating Factors

We examine the average ratings for each student who participated, as shown in Table 4. We find that for 3 of the 4 students, the raters perceived more rapport in the exchanges where Quinn adapted by the shift+contour adaptation than in any other condition. Listening to these recordings, we identify an imbalance in terms of content spoken. In most scenarios, Quinn and the student engaged in conversation like the transcript in Section 5.1. In other cases, Quinn would introduce an off-topic statement. For example,

Q: This is not very fun, are we almost done?
S: Math can be fun! But yeah…we're almost done

		Average Rapport				Average Naturalness			
	gender	control	mirror	s-contour	s-flatten	control	mirror	s-contour	s-flatten
Student 1	F	3.61	3.68	**3.74**	3.38	2.14	1.69	**2.41**	1.74
Student 2	F	3.58	3.70	**3.78**	3.59	2.36	1.94	**2.39**	1.93
Student 3	M	3.64	3.65	**3.80**	3.55	**2.28**	1.74	2.27	1.83
Student 4	M	**3.79**	3.36	3.29	3.25	**2.50**	1.87	2.05	1.68

Table 4: Descriptive statistics for each student; bold values indicate the highest rapport/naturalness score for that student

There is support in past work that off-task social conversation increases rapport, particularly in educational dialogues [25]. Given the possibility the raters are considering the content of exchanges in their evaluations of rapport, we annotate the exchanges as either social (off-topic and not about the problem), or *not* social (on-topic and about the problem). In addition, we consider that we designed Quinn to entrain to the previous turn made by the student. In the exchanges rated, we counter balanced between exchanges where the rater would hear Quinn speak first and scenarios where the rater would hear the student speak first. In the latter, the turn to which Quinn is adapting is audible. We suspect the raters are perceiving more differences in the rapport produced when they can hear the speech to which Quinn is adapting.

5.3.3 Statistical Analysis Incorporating Speaking Order and Social Context

To explore the effect of the social exchanges versus non-social exchanges as well as the order in which Quinn speaks, we run a 3-way ANOVA with rapport as the independent variable, including the type of adaptation, whether Quinn speaks first or second, and the social/not-social annotations as factors. The ANOVA analysis indicates statistically significant interactions between all combinations of factors except for the highest order interaction (all 3 factors). F-scores and p-values are shown in Table 5.

Table 5: 3-way ANOVA with rapport as dependent variable

Factor	F-Score	*p*
Type of Adaptation	6.8	< 0.001
Social/Non-Social Exchange	1.3	0.25
Quinn Speaks First/Second	3.6	0.06
Adaptation x Social Exchange	6.1	< 0.001
Adaptation x Quinn Speaking	7.5	< 0.001
Social Exchange x Quinn Speaking	12.7	< 0.001
3-Way Interaction	2.0	0.11

Finding significant 2-way interactions for all combinations of factors, we run pairwise comparisons for further analysis. In social exchanges, the type of adaptation results in significantly different levels of rapport. When Quinn speaks second, shift+contour has significantly higher rapport than the control ($p = 0.03$) and shift+flatten ($p < 0.001$). The difference with mirror partner is nearly significant ($p = 0.08$). Pitch adaptation in non-social exchanges or when Quinn speaks first appear to have less of an effect.

6. DISCUSSION & CONCLUSION

In reviewing the results above, we find that in terms of rapport produced, differences between the pitch adaptations become the most notable when we incorporate social/non-social annotations. We find that in social exchanges, shift-contour produces significantly more rapport than the other adaptations and the control. This suggests that in future work, we should consider an adaptation such as shift-contour and that it may be prudent to pick and choose when an agent entrains. This is supported by prior work on rapport and entrainment which shows that off-task social conversation increases rapport and that humans entrain more in off-task, social dialogues [21].

While we do find support for our hypothesis that shift+flatten produces the least rapport, our hypothesis regarding mirror partner producing more rapport than shift+contour or the control is not supported by the results. Listening to the exchanges, this result is mostly likely due to our original concern that mirror partner results in mismatches between pitch contour and syntactic and semantic structure. This is supported by the finding that mirror partner is significantly less natural. Considering mirror partner did receive very low naturalness scores, the rapport perceived for this adaptation is relatively high. This suggests that overcoming the issues with syntactic and semantic structure with a more nuanced adaptation accounting for contextual dependencies is necessary if we wish to explore mirror partner in the future.

We conclude that adapting to the speaker does appear to be have an effect on naturalness and rapport and we find that shifting the contour by pitch mean is one form of adaptation we can accomplish in a manner which sounds as natural as current text-to-speech technologies while significantly increasing perceptions of rapport. Future work includes extending these findings and running a more extensive analysis focusing specifically on the effects of adaptation on pitch mean in regards to learning with a learning companion and assessing effects on self-reported rapport. We also intend to explore additional acoustic-prosodic features for adaptation and additional adaptation models where we can incorporate more refined techniques based on phoneme length and pauses as well as other forms of entrainment, such as convergence, where individuals adapt over the course of a conversation.

10. REFERENCES

[1] Borrie, S. A., & Liss, J. M. Rhythm as a coordinating device: entrainment with disordered speech. *Journal of Speech, Language, and Hearing Research*, *57*(3), 815-824, 2014.

[2] Boersma, Paul & Weenink, David. Praat: doing phonetics by computer [Computer program]. Version 5.4.12, retrieved 10 July 2015 from http://www.praat.org/

[3] Buhrmester, M., Kwang, T., & Gosling, S. D. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?. *Perspectives on psychological science*, *6*(1), 3-5, 2011.

[4] Chou, C. Y., Chan, T. W., & Lin, C. J. Redefining the learning companion: the past, present, and future of educational agents. *Computers & Education*, 40(3), 255-269. 2003.

[5] Coulston, R., Oviatt, S., & Darves, C. "Amplitude convergence in children's conversational speech with animated personas." *Proceedings of the 7th International Conference on Spoken Language Processing*. Vol. 4. 2002.

[6] Eide, Ellen M., and Robert E. Donovan. "Methods for generating pitch and duration contours in a text to speech system." U.S. Patent No. 6,101,470, 2000.

[7] Frager, S., & Stern, C. Learning by teaching. *The Reading Teacher*, 403-417, 1970.

[8] Friedberg, H., Litman, D., & Paletz, S. B. Lexical entrainment and success in student engineering groups. In *Spoken Language Technology Workshop (SLT), 2012 IEEE* (pp. 404-409). IEEE, 2012.

[9] Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. "Creating rapport with virtual agents." In *Intelligent Virtual Agents*, Springer, 125-138, 2007.

[10] Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 181-188). ACM, 2013.

[11] Iio, T., Shiomi, M., Shinozawa, K., Takaaki, A., Hagita, N., & Shimohara, K. Entrainment between speech and gestures in human-robot interaction. In *SICE Annual Conference 2010, Proceedings of* (pp. 2769-2774). IEEE, 2010.

[12] ITU-T Recommendation P.85. Telephone transmission quality subjective opinion tests. A method for subjective performance assessment of the quality of speech voice output devices, 1994.

[13] Jurcıcek, F., Keizer, S., Gašic, M., Mairesse, F., Thomson, B., Yu, K., & Young, S. "Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk." Proceedings of INTERSPEECH. Vol. 11. 2011.

[14] Kumar, R., Ai, H., Beuth, J. L., & Ros é, C. P. Socially capable conversational tutors can be effective in collaborative learning situations. In *Intelligent tutoring systems* (pp. 156-164). Springer Berlin Heidelberg, 2010.

[15] Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, *27*(3), 145-162, 2003.

[16] Lee, C. C., Black, M., Katsamanis, A., Lammert, A. C., Baucom, B. R., Christensen, A. & Narayanan, S. S. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *INTERSPEECH* (pp. 793-796), 2010.

[17] Levitan, R., Gravano, A., Willson, L., Benus, S., Hirschberg, J., & Nenkova, A. "Acoustic-prosodic entrainment and social behavior." In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*. Association for Computational Linguistics, 2012.

[18] Levitan, R. *Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue* (Doctoral dissertation, Columbia University), 2014.

[19] Lubold, N., & Pon-Barry, H. "Acoustic-Prosodic Entrainment and Rapport in Collaborative Learning Dialogues". In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, ACM, 5-12, 2014.

[20] Lubold, N., & Pon-Barry, H. "A comparison of acoustic-prosodic entrainment in face-to-face and remote collaborative learning dialogues." In *Spoken Language Technology Workshop (SLT), 2014 IEEE* (pp. 288-293). IEEE, 2014.

[21] Lubold, N., Walker, E., & Pon-Barry, H. "Relating Entrainment, Grounding, and Topic of Discussion in Collaborative Learning Dialogues." In *Proceedings of Computer Supported Collaborative Learning*, 2015.

[22] Reeves, B., & Nass, C. *How people treat computers, television, and new media like real people and places* (p. 119). CSLI Publications and Cambridge university press, 1996.

[23] Nenkova, A., Gravano, A., & Hirschberg, J. "High frequency word entrainment in spoken dialogue." In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008.

[24] Obayashi, F., Shimoda, H., & Yoshikawa, H. Construction and evaluation of CAI system based on learning by teaching to virtual student. In *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics*. Vol. 3, 94-99, 2000.

[25] Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N., & Cassell, J. “Oh dear Stacy! Social Interaction, Elaboration, and Learning with Teachable Agents.” In Proceedings of the *SIGCHI Conference on Human Factors in Computing Systems* (pp. 39-48). ACM, 2012.

[26] Reitter, D., Keller, F., & Moore, J. D. A computational cognitive model of syntactic priming. *Cognitive science*, *35*(4), 587-637, 2011.

[27] Stylianou, Y. Applying the harmonic plus noise model in concatenative speech synthesis. *Speech and Audio Processing, IEEE Transactions on*, *9*(1), 21-29, 2001.

[28] Suzuki, N., Takeuchi, Y., Ishii, K., & Okada, M. Effects of echoic mimicry using hummed sounds on human-computer interaction. Speech Communication, 40(4), 559–573, 2003.

[29] Suzuki, N., and Katagiri, Y. "Prosodic alignment in human–computer interaction." *Connection Science* 19.2, 131-141, 2007.

[30] Thomason, J., Nguyen, H. V., & Litman, D. “Prosodic entrainment and tutoring dialogue success.” In *Artificial Intelligence in Education*. Springer Berlin Heidelberg, 2013.

[31] Violante, L., Zivic, P. R., & Gravano, A. “Improving speech synthesis quality by reducing pitch peaks in the source recordings”. In *HLT-NAACL*, 502-506, 2013.

[32] Vygotsky, L. S. Mind and society: The development of higher mental processes, 1978.

[33] Watanabe, Tomio. "Effects of pitch adaptation in prosody on human-machine verbal communication." *Advances in Human Factors/Ergonomic*, 20, 269-274, 1995.

[34] Wallace, R. (2003). The elements of AIML style. *Alice AI Foundation*.