Towards Improving the Naturalness of Social Conversations with Dialogue Systems

Matthew Marge, João Miranda, Alan W Black, Alexander I. Rudnicky
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{mrmarge,jmiranda,awb,air}@cs.cmu.edu

Abstract

We describe an approach to improving the naturalness of a social dialogue system, Talkie, by adding disfluencies and other content-independent enhancements to synthesized conversations. We investigated whether listeners perceive conversations with these improvements as natural (i.e., human-like) as human-human conversations. We also assessed their ability to correctly identify these conversations as between humans or computers. We find that these enhancements can improve the perceived naturalness of conversations for observers "overhearing" the dialogues.

1 Introduction

An enduring problem in spoken dialogue systems research is how to make conversations between humans and computers approach the naturalness of human-human conversations. Although this has been addressed in several goal-oriented dialogue systems (e.g., for tutoring, question answering, etc.), *social* dialogue systems (i.e., non-task-oriented) have not significantly advanced beyond so-called "chatbots". Proper social dialogue systems (Bickmore and Cassell, 2004; Higuchi et al., 2002) would be able to conduct open conversations, without being restricted to particular domains. Such systems would find use in many environments (e.g., human-robot interaction, entertainment technology).

This paper presents an approach to improving a social dialogue system capable of chatting about the news by adding content-independent enhancements to speech. We hypothesize that enhancements such as explicit acknowledgments (e.g., *right*, *so*, *well*) and disfluencies can make human-computer conversations sound indistinguishable from those between two humans.

Enhancements to synthesized speech have been found to influence perception of a synthetic voice's hesitation (Carlson et al., 2006) and personality (Nass and Lee, 2001). Andersson et al. (2010) used machine learning techniques to determine where to include conversational phenomena to improve synthesized speech. Adell et al. (2007) developed methods for inserting filled pauses into synthesized speech that listeners found more natural. In these studies, human judges compared utterances in isolation with and without improvements. In our study, we focus on a holistic evaluation of naturalness in dialogues and ask observers to directly assess the naturalness of conversations that they "overhear".

2 The Talkie System

Talkie is a spoken dialogue system capable of having open conversations about recent topics in the news. This system was developed for a dialogue systems course (Lim et al., 2009). Interaction is intended to be unstructured and free-flowing, much like social conversations. Talkie initiates a conversation by mentioning a recent news headline and invites the user to comment on it.

The system uses a database of news topics and human-written comments from the "most blogged about articles" of the New York Times (NYT)[1]. Comments are divided into single sentences to approximate the length of a spoken response. Given a user's utterance (e.g., keywords related to the topic), Talkie responds with the comment that most closely resembles that utterance. Talkie may access any comment related to the topic under discussion (without repetition). The user may choose to switch to a different topic at any time (at which point Talkie will propose a different topic from its set).

[1] http://www.nytimes.com/gst/mostblogged.html
Follow links to each article's comment section.

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 91–94,
The University of Tokyo, September 24-25, 2010.

3 Study

We performed a study to determine if the perceived naturalness of conversations could be improved by using heuristic enhancements to speech output. Participants "overheard" conversations (similar to Walker et al. (2004)). Originally typed interactions, the conversations were later synthesized into speech using the Flite speech synthesis engine (Black and Lenzo, 2001). For distinctiveness, conversations were between one male voice (rms) and one female voice (slt). The voices were generated using the CLUSTERGEN statistical parametric synthesizer (Black, 2006). All conversations began with the female voice.

3.1 Dialogue Content

We considered four different conversation types: (1 & 2) between a human and Talkie (human-computer and computer-human depending on the first speaker), (3) between two humans on a topic in Talkie's database (human-human), and (4) between two instances of Talkie (computer-computer). The human-computer and computer-human conditions differed from each other by one utterance; that is, one was a shifted version of the other by one dialogue turn. The human-computer conversations were collected from two people (one native English speaker, one native Portuguese speaker) interacting with Talkie on separate occasions. For human-human conversations, Talkie proposed a topic for discussion. Each conversation contained ten turns of dialogue. To remove any potential effects from the start and end content of the conversations, we selected the middle three turns for synthesis. Each conversation type had five conversations, each about one of five recent headlines (as of May 2010).

3.2 Heuristic Enhancements

We defined a set of rules that added phenomena observed in human-human spoken conversations. These included filled pauses, word repetitions, silences, and explicit acknowledgments. Conversations in this study were enhanced manually by following the set of rules described in Figure 1; an example is shown in Figure 2.

3.3 Participants and Task

Eighty participants were recruited from Amazon's Mechanical Turk[2] (MTurk) for this between-subjects study. Only workers with a 95% HIT (human intelligence task) approval rating could participate. There were three experimental dimensions that we counterbalanced across participants: the four conversation "types" described in Section 3.1, the order of the five topics, and the presence or absence of heuristic enhancements to the synthesized speech that participants heard. Each participant listened to every topic such that exactly five participants heard each condition.

Participants made two judgments per conversation: rating the "naturalness" of a conversation based on a four-point Likert scale and ranking the conversant identities (i.e., as two humans, two computers, or a computer and a human) from most to least confident. We used the non-parametric Kruskal-Wallis one-way analysis of variance to determine statistically significant differences in naturalness responses across groups (naturalness being the dependent measure).

Category I - *Explicit Acknolwedgements*

- inserted sparingly at the beginning of sentences when grammatical (e.g., *well*, *so*, *you know*, *right*).

Category II - *Filled pauses / repetitions*

- no more than three per dialogue
- no more than two in any one sentence
- placed repetitions in places where speaker may have hesitation (e.g., after "I think...")
- utterances may have a filled pause or a repetition followed by a filled pause.

Category III - *Silences*

- added sparingly after filled pauses.

Figure 1: Heuristics used for adding enhancements to synthesized dialogue.

4 Results

We analyzed two measures: (1) naturalness ratings of conversations, and (2) frequencies from highest-rank to lowest-rank of who the participant believed to be the speakers in each conversation. We found no statistically significant ordering effects or biases due to the topic of conversation.

4.1 Speaker Type Condition Results

Results for naturalness scores are summarized in Figure 3. As expected, mean naturalness was highest for the human-human condition and lowest for the computer-computer condition. We found

[2] http://www.mturk.com

Human-computer conversation (original, **enhancements)**

h: Well, that sounds ridiculous, you get taxed proportionally to your paycheck.
c: So, let them retire on social security payments, then you will see changes with <**uh**> **with** an overabundance of funds.
h: I don't see how this would be possible.
c: You know, you don't have the money yet, and you have just increased your debt by 2.5 trillion dollars.
h: Exactly why should we be raising taxes on the rich?
c: With our current politics of division, I <**umm**> worry we have the will or determination to address these issues.

Figure 2: Example conversation with heuristic enhancements marked in bold.

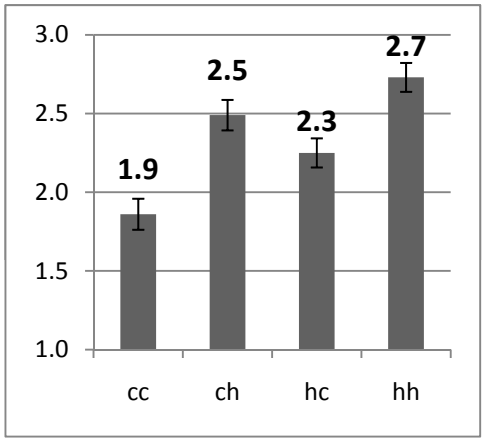

Figure 3: Naturalness across the speaker type condition.

no statistically significant difference in naturalness ratings for the computer-human condition compared to the human-computer condition (H(1) = 2.94; p = 0.09). Also, the computer-computer condition was significantly different from all other conditions, suggesting that conversation flow is an important factor in determining the naturalness of a conversation (H(3) = 42.49, $p < 0.05$).

People rated conversations involving a computer and a human similarly to human-human conversations (without enhancements). There were no statistically significant differences between the three conditions *cc*, *ch*, and *hc* (H(2) = 5.36, p = 0.06). However, a trend indicated that *hc* naturalness ratings differed from those of the *ch* and *hh* conditions. Conversations from the *hc* condition had much lower (18%) mean naturalness ratings compared to their *ch* counterparts, even though they were nearly equivalent in content.

4.2 Heuristic Enhancements Results

There were significant differences in naturalness ratings when heuristic enhancements were present (H(1) = 17.49, $p < 0.05$). Figure 4 shows that the perceived naturalness was on average higher with heuristic enhancements. Overall, mean naturalness improved by 20%. This result agrees with findings from Andersson et al. (2010).

Computer-computer conversations had the highest relative improvement (42%) in mean naturalness. Naturalness ratings were significantly different when comparing these conversations with and without enhancements (H(1) = 11.77, $p < 0.05$). Content-free conversational phenomena appear to compensate for the lack of logical flow in these conversations. According to Figure 5, after enhancements people are no better than chance at correctly determining the speakers in a computer-computer conversation. Thus the heuristic enhancements clearly affect naturalness judgments.

Even the naturalness of conversations with good logical flow can improve with heuristic adjustments; there was a 26% relative improvement in the mean naturalness of human-human conversations. Participant ratings of naturalness were again significantly different (H(1) = 12.45, $p < 0.05$). Note that these conversations were originally typed dialogue. As such, they did not capture turn-taking properties present in conversational speech. When enhanced with conversational phenomena, they more closely resembled natural spoken conversations. As shown in Figure 5, people are more likely than chance to correctly identify two humans as being the participants in the dialogue after these enhancements were applied to speech.

Conversations with one computer and one human also benefited from heuristic enhancements. Improvements in naturalness were marginal, however. Naturalness scores in the *hc* condition improved by 16%, but this improvement was only a trend (H(1) = 3.66, p = 0.06). Improvement was negligible in the *ch* condition. Participants selected the correct speakers in human-computer dialogues no better than random. We note that participants tended to avoid ranking conversations as "human & computer" with confidence (i.e., the highest rank). A significant majority (267 out of 400) of second-rank selections were "human & computer." Participants tended to order conditions

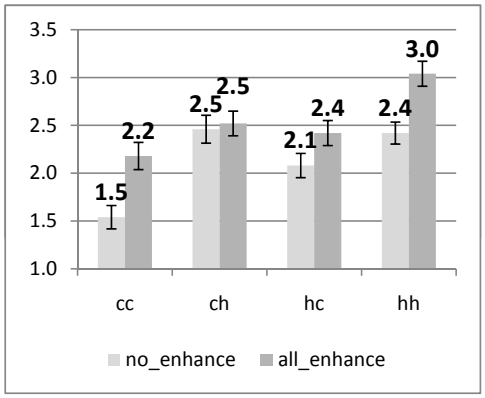

Figure 4: Mean naturalness across enhancement conditions.

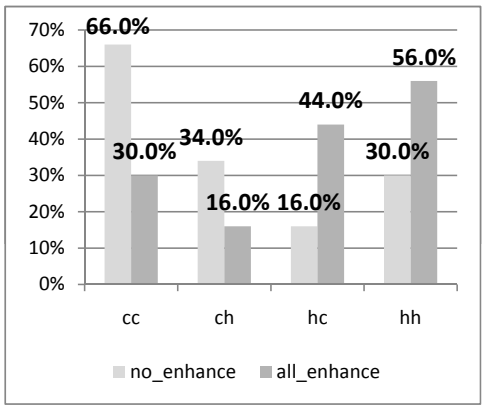

Figure 5: Percentage of participants' selections of members of the conversation that were correct.

from all human to all computer or vice-versa.

5 Conclusions

We have shown that content-independent heuristics can be used to improve the perceived naturalness of conversations. Our conversations sampled a variety of interactions using Talkie, a social dialogue system that converses about recent news headlines. An experiment examined the factors that could influence how external judges rate the naturalness of these conversations.

We found that without enhancements, people rated conversations involving a human and a computer similarly to conversations involving two humans. Adding heuristic enhancements produced different results, depending on the conversation type: computer-computer and human-human conversations had the best gain in naturalness scores. Though it remains to be seen if people are always influenced by such enhancements, they are clearly useful for improving the naturalness of human-computer dialogues.

Future work will involve developing methods to automatically inject enhancements into the synthesized speech output produced by Talkie, as well as determining whether other types of systems can benefit from these techniques.

Acknowledgments

We would like to thank Aasish Pappu, Jose-Pablo Gonzales Brenes, Long Qin, and Daniel Lim for developing the Talkie dialogue system.

References

J. Adell, A. Bonafonte, and D. Escudero. *Filled pauses in speech synthesis: Towards conversational speech.* In TSD'07, Pilsen, Czech Republic, 2007.

S. Andersson, K. Georgila, D. Traum, M. Aylett, and R.A.J. Clark. *Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection.* In the 5th International Conference on Speech Prosody, Chicago, Illinois, USA, 2010.

T. Bickmore and J. Cassell. *Social Dialogue with Embodied Conversational Agents.* J. van Kuppevelt, L. Dybkjaer, and N. Bernsen (eds.), Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems. New York: Kluwer Academic.

A. Black. CLUSTERGEN*: A Statistical Parametric Synthesizer using Trajectory Modeling.* In Interspeech'06 - ICSLP, Pittsburgh, PA, 2006.

A. Black and K. Lenzo. *Flite: a small fast run-time synthesis engine.* In ISCA 4th Speech Synthesis Workshop, Scotland, 2001.

R. Carlson and K. Gustafson and E. Strangert. *Cues for Hesitation in Speech Synthesis.* In Interspeech'06 - ICSLP, Pittsburgh, PA, 2006.

S. Higuchi, R. Rzepka, and K. Araki. *A casual conversation system using modality and word associations retrieved from the web.* In EMNLP'08. Honolulu, Hawaii, 2008.

D. Lim, A. Pappu, J. Gonzales-Brenes, and L. Qin. *The Talkie Spoken Dialogue System.* Unpublished manuscript, Carnegie Mellon Univeristy, 2009.

C. Nass and K. M. Lee. *Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction.* Journal of Experimental Psychology: Applied 7 (2001) 171-181.

M. A. Walker, S. J. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, G. Vasireddy. *Generation and evaluation of user tailored responses in multimodal dialogue.* Cognitive Sci. 28 (2004) 811-840.