Using emotion to gain rapport in a spoken dialog system

Jaime C. Acosta
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
jcacosta@miners.utep.edu

Abstract

This paper describes research on automatically building rapport. This is done by adapting responses in a spoken dialog system to users' emotions as inferred from nonverbal voice properties. Emotions and their acoustic correlates will be extracted from a persuasive dialog corpus and will be used to implement an emotionally intelligent dialog system; one that can recognize emotion, choose an optimal strategy for gaining rapport, and render a response that contains appropriate emotion, both lexically and auditory. In order to determine the value of emotion modeling for gaining rapport in a spoken dialog system, the final implementation will be evaluated using different configurations through a user study.

1 Introduction

As information sources become richer and technology advances, the use of computers to deliver information is increasing. In particular, interactive voice technology for information delivery is becoming more common due to improvements in technologies such as automatic speech recognition, and speech synthesis.

Several problems exist in these voice technologies including speech recognition accuracy and lack of common sense and basic knowledge. Among these problems is the inability to achieve rapport.

Gratch *et al.* (2007) defines rapport as *a feeling of connectedness that seems to arise from rapid and contingent positive feedback between partners and is often associated with socio-emotional processes*. In the field of neuro-linguistics, O'Connel and Seymour (1990) stated that matching or complimenting voice features such as volume, speed, and intonation, is important to gain rapport. Shepard et al.'s Communication Accommodation Theory (2001) states that humans use prosody and backchannels in order to adjust social distance with an interlocutor. These features of voice can also be associated with emotions.

Previous work has shown that automated systems can gain rapport by reacting to user gestural nonverbal behavior (Chartrand and Bargh, 1999; Gratch et al., 2007; Cassell and Bickmore, 2003). In contrast, this research looks at how rapport can be gained through voice-only interaction.

Preliminary analysis of human-human dialog provides evidence that shifts in pitch, associated with emotion by two judges, are used by an interlocutor for persuasion. Figure 1 shows the pitch of a sound snippet from the corpus and how it differs from neutral, computer synthesized voice (produced using MaryTTS). This illustrates the more general fact that when humans speak to each other, we display a variety of nonverbal behaviors in voice, especially when trying to build rapport. The main hypothesis of this research is that a spoken dialog system with emotional intelligence will be effective for gaining rapport with human users.

The rest of this paper is structured as follows: first, related work is reviewed and current limitations for building automated rapport are described. Afterwards, the hypotheses and expected contributions of this work are described along with the research approach. Lastly, broader significance of this work is discussed.

Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium, pages 49–54,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

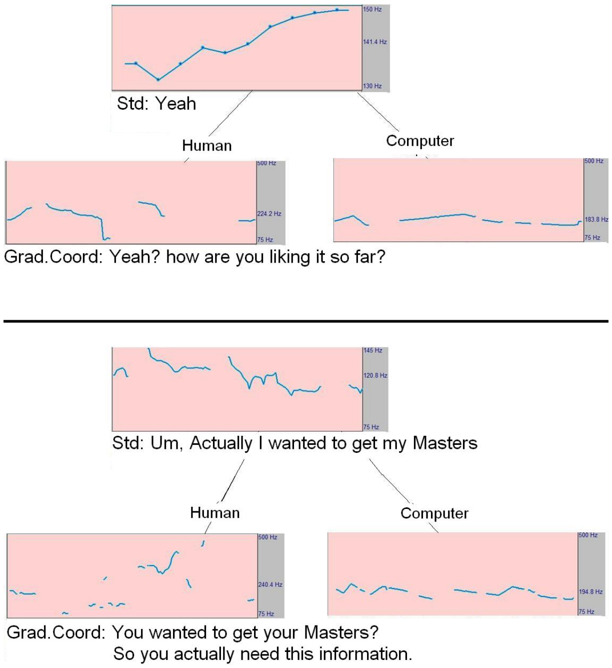

Figure 1: Pitch levels of a conversation taken from the persuasive dialog corpus includes a student (Std) and a graduate coordinator (Grad.Coord). Pitch was analyzed using the Praat software. It can be seen that the student displays rich prosody in voice (tree parents) and that the human response (left branch) contains more varied prosody than the computer synthesized voice (right branch).

2 Related Work

Communication Accommodation Theory states that people use nonverbal feedback to establish social distance during conversation. In order to gain rapport, people would most likely want to decrease social distance in order to achieve the connectedness and smoothness in conversation that is seen in human social interaction. Research in human-computer interaction has pursued these nonverbal behaviors through appropriate backchanneling, head nods, and gaze techniques, but still missing is attention to user emotional state, which can be detected through some of these nonverbal behaviors in voice.

Two methods for describing emotions are discrete and dimensional. Discrete emotions include anger, disgust, fear, joy, sadness, and surprise. Dimensional emotions use two or more components to describe affective state. More commonly used dimensions are Osgood *et al.*'s (1957) evaluation (a.k.a. valence), activity, and potency (a.k.a. power). Emotion research has had limited success at detecting discrete emotions, e.g. (D'Mello et al., 2008). In the tutoring domain, some have looked at appropriately responding to students based on their prosody in voice (Hollingsed and Ward, 2007). The difficulty of recognizing discrete emotions exists because humans typically show more subtle emotions in most real human-human interactions (Batliner et al., 2000). Forbes *et al.* (2004) had promising results by looking at a three-class set of emotions (positive, negative, neutral).

The intent of this research is to develop a method for detecting three dimensions of emotion from voice in order to build rapport. There is a possibility that using a dimensional approach will enable more accurate modeling of subtle emotions that exist in spontaneous human-human dialogs.

3 Hypotheses and Expected Contributions

The main hypothesis of this work is that a spoken dialog system with emotional intelligence will be more effective for gaining rapport than a spoken dialog system without emotional intelligence. In order to test this hypothesis, I will implement and evaluate a spoken dialog system. This system will choose topics and content depending on user emotional state. The resulting system will advance the state of the art in three technologies: recognizing appropriate emotion, planning accordingly, and synthesizing appropriate emotion. The system will also demonstrate how to integrate these components.

In addition to choosing the correct content based on user emotional state, this research will investigate the effect of adding emotion to voice for rapport. The second hypothesis of the research is that expressing emotion in voice and choosing words, compared to expressing emotion only by choosing words, will be more effective for building rapport with users.

4 Approach

This section outlines the steps that have been completed and those that are still pending to accomplish the goals of the research.

4.1 Corpus Analysis and Baseline System

This work is based on a persuasive dialog corpus consisting of audio recordings of 10 interactions averaging 16 minutes in length. The corpus consists of rougly 1000 turns between a graduate coordinator and individual students. The graduate coordinator was a personable female staff member who was hired by the University to raise the graduate student count. The students were enrolled in an introductory Computer Science course and participated in the study as part of a research credit required for course completion. The students had little knowledge of the nature or value of graduate school and of the application process. Preliminary analysis of the corpus showed evidence of a graduate coordinator building rapport with students by using emotion.

A baseline system built using commercial state-of-the-art software was implemented based on the corpus (mainly the topics covered). Informal user comments about the baseline system helped determine missing features for automated rapport building technology. One salient feature that is missing is attention to emotion in voice. This confirmed the direction of this research.

This corpus was transcribed and annotated with dimensional emotions (activation, valence, and power) by two judges. Activation is defined as sounding ready to take action, valence is the amount of positive or negative sound in voice, and power is measured by the amount of dominance in voice. The dimensions are annotated numerically on scales from -100 to +100.

The following are examples taken from the corpus with annotated acoustic features.

- Example 1
  Grad.Coord(GC1): *So you're in the 1401 class?* [rising pitch]

  Subject(S1): *Yeah.* [higher pitch]

  GC2: *Yeah? How are you liking it so far?* [falling pitch]

  S2: *Um, it's alright, it's just the labs are kind of difficult sometimes, they can, they give like long stuff.* [slower speed]

  GC3: *Mm. Are the TAs helping you?*
  [lower pitch and slower speed]

  S3: *Yeah.* [rising pitch]

  GC4: *Yeah.* [rising pitch]

  S4: *They're doing a good job.*
  [normal pitch and normal speed]

  GC5: *Good, that's good, that's good.*
  [normal pitch and normal speed]

- Example 2
  GC6: *You're taking your first CS class huh.*
  [slightly faster voice]

  S5: *Yeah, I barely started.* [faster voice]

  GC7: *How are you liking it?*
  [faster voice, higher pitch]

  S6: *Uh, I like it a lot, actually, it's probably my favorite class.* [faster, louder]

  GC8: *Oh good.* [slower, softer]

S7: *That I'm taking right now yeah.* [slightly faster, softer]

GC9: *Oh that's good. That's exciting.* [slow and soft then fast and loud]

GC10: *Then you picked the right major you're not gonna change it three times like I did.* [faster, louder]

In the first example, the coordinator noticably raises her pitch at the end of her utterance. This is probably so that she can sound polite or interested. On line S2, the subject displays a falling pitch (which sounds negative) and the coordinator responds with a lower fundamental frequency and a slower speed. The subject sounds unsure by displaying a rising pitch in his answer (S3). The coordinator mirrors his response (GC4) and finally both interlocutors end with normal pitch and normal speed.

In the second example, the subject speaks faster than usual (S5). The coordinator compensates by adjusting her speed as well. From S6 through GC8, when the subject's voice gets louder, the coordinator's voice gets softer, almost as though she is backing off and letting the subject have some space. In GC9 the coordinator responds to the student's positive response (liking the class) and becomes immediately faster and louder.

A next step for the analysis is to determine the most expressive acoustic correlates for emotions. Informal auditory comparisons show some possible correlations (see Table 1). These correlations seem promising because many correspond with previous work (Schroder, 2004).

Table 1: Informal analysis reveals acoustic correlates possibly associated with the dimensions of emotion

Dimension	High	Low
Activeness	Faster, more varied pitch, louder	Slower, less varied pitch, softer
Valence	Higher pitch throughout, laughter, speed up	Falling ending pitch, articulation of words, increasing loudness
Power	Faster, louder, falling ending pitch, articulation of word beginnings, longer vowels	Softer, higher pitch throughout, quick rise in pitch, smoother word connection

The emotion annotations of the two judges show that strategies for adaptive emotion responses can be extracted from the corpus. Communication Accomodation Theory states that interlocutors mirror nonverbal behaviors during interaction when attempting to decrease social distance. The coordinator's emotional responses were correlated with the student's emotional utterances to determine if emotional mirroring (matching student emotion and coordinator response) was present in the persuasive dialog corpus. This was the case in the valence dimension, which showed a correlation coefficient of 0.34. However, regarding power, there was an inverse relationship; if the student showed more power, the coordinator showed less (–0.30 correlation coefficient). Activation showed a small correlation coefficient (–0.14).

To realize a spoken dialog system that could model this responsive behavior, machine learning was used. The students' three emotion dimensions were taken as attributes and were used to predict the coordinators emotional responses using Bagging with REPTrees. Measuring the correlations between the predictions of the model and the actual values in the corpus revealed correlation coefficients of 0.347, 0.344, and 0.187 when predicting the coordinator's valence, power, and activation levels, respectively.

4.2 Full System

The full system will provide a means to evaluate whether emotion contributes to automated rapport building. This system will be based on several available technologies and previous research in spoken dialog systems.

Figure 2 shows the different components anticipated for the full system. The components that will be implemented for this research include emotion recognition, user modeling components, and text and emotion strategy databases. The other components will be based on available open source software packages. The implementation effort also in-

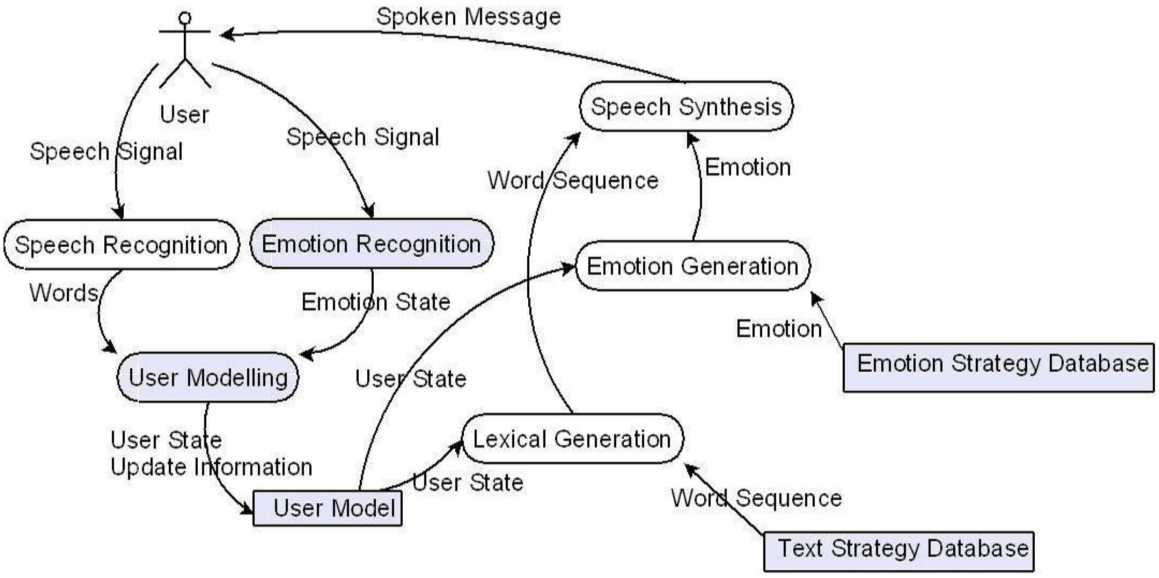

Figure 2: Full System Dataflow Diagram

cludes the integration of all components.

The following is a scenario that depicts how the full system will operate.

1. The system begins by saying "How are you doing today?"

2. The user says "I'm doing good" with a negative sounding voice.

3. The voice signal is then processed through the speech recognizer and emotion recognizer in parallel. The speech recognizer extracts words from the voice signal while the emotion recognizer extracts emotion.

4. This data is sent to the user modeling component which determines the immediate user state based only on the current emotion and the words spoken. In this scenario, the user's state will be negative even though the user stated otherwise.

5. This user state update information is then passed to the user model which updates the current user state. This component contains knowledge, beliefs and feelings of the user. Since there was no previous user state, the current emotion is set to negative. Stored in user knowledge will be the fact that the user was asked "How are you doing today?". Some information about the user's contradictory state is stored as user beliefs: stated good, but sounds negative.

6. Next, this information is used to select some predefined text from the lexical generation along with an associated emotion from the emotion strategy database (these two are done in parallel). Since the user's state is negative, the system may choose to ask another question such as "ok, do you have any concerns?" with a negative sounding voice (to mirror the valence dimension). In contrast, if the user was positive, the system may have chosen something similar to "great, let's get going then" with a highly positive voice.

7. Lastly, the text with corresponding emotion coloring is rendered to speech and played to the user by the speech synthesis component.

4.3 Evaluation

To achieve the final goal of determining whether emotion helps gain rapport, the final system described herein will be evaluated.

The final system will be configurable; it will allow for enabling emotion in voice (*voiced*) or disabling the emotions in voice (*not voiced*). In addition, there will be a control configuration, perhaps one that will display a random emotion (*random*). A user study (hopefully within subjects) will be conducted that will ask users to interact with four versions of the system (baseline, *voiced*, *not voiced*, and *random*). A post-test questionnaire consisting of Likert scales will ask users how much rapport they felt with each version of the system. In addition, some objective metrics such as disfluency count and interaction time will be collected. This will help test the two hypotheses of this research. First, it is expected that subjects will have more rapport with the *not voiced* configuration than with the baseline system. The second hypothesis will be verified by determining if subjects have more rapport with the *voiced* than with the *not voiced* system. The *random* configuration will be used to determine whether the system's adaptive responses are better than random responses.

5 Broader Significance

This research addresses methods for gaining rapport as an important dimension of successful human-computer interaction, and one likely to be useful even for business-like dialogs. For example, building rapport with customers can decrease the number of disfluencies, which are currently a problem for speech recognizers. In addition, customer support systems will have the ability to tailor responses to decrease negative emotion.

Similarly, the learned rules for detecting emotion and responding appropriately could be used to train people how to more effectively gain rapport. Lastly, this work can supplement other rapport research that uses other forms of nonverbal behavior such as gaze and gestures seen especially in embodied conversational agents.

6 Acknowledgements

I would like to thank my advisor, Nigel Ward for his help. Also, I would like to thank Anais Rivera and Sue Walker for the collection of the persuasive dialog corpus and Jun Zheng for his help in fine tuning the baseline system. This work is supported in part by NSF grant IIS-0415150.

References

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. 2000. Desperately Seeking Emotions or: Actors, Wizards, and Human Beings. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. ISCA.

J. Cassell and T. Bickmore. 2003. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and User-Adapted Interaction*, 13(1):89–132.

T.L. Chartrand and J.A. Bargh. 1999. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910.

S.K. D'Mello, S.D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser. 2008. Automatic detection of learners affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1):45–80.

K. Forbes-Riley and D. Litman. 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. *Proc. Human Language Technology Conf. of the North American Chap. of the Assoc. for Computational Linguistics (HLT/NAACL).*

J. Gratch, N. Wang, A. Okhmatovskaia, F. Lamothe, M. Morales, R.J. van der Werf, and L. Morency. 2007. Can Virtual Humans Be More Engaging Than Real Ones? *12th International Conference on Human-Computer Interaction.*

Tasha K. Hollingsed and Nigel G. Ward. 2007. A combined method for discovering short-term affect-based response rules for spoken tutorial dialog. *Workshop on Speech and Language Technology in Education (SLaTE).*

J. O'Connor and J. Seymour. 1990. *Introducing neurolinguistic programming*. Mandala.

C.E. Osgood. 1957. *The Measurement of Meaning*. University of Illinois Press.

M. Schroder. 2004. Dimensional Emotion Representation as a Basis for Speech Synthesis with Non-extreme Emotions. *In Proceedings Workshop Affective Dialogue Systems*, 3068:209–220.

C.A. Shepard, H. Giles, and B.A. Le Poire. 2001. Communication accommodation theory. *The new handbook of language and social psychology*, pages 33–56.