INTERSPEECH 2009 BRIGHTON

Responding to User Emotional State by Adding Emotional Coloring to Utterances

Jaime C. Acosta and Nigel G. Ward

Department of Computer Science, University of Texas at El Paso, USA

jaimeacosta@acm.org, nigelward@acm.org

Abstract

When people speak to each other, they share a rich set of non-verbal behaviors such as varying prosody in voice. These behaviors, sometimes interpreted as demonstrations of emotions, call for appropriate responses, but today's spoken dialog systems lack the ability to do so. We collected a corpus of persuasive dialogs, specifically conversations about graduate school between a staff member and students, and had judges label all utterances with triples indicating the perceived emotions, using the three dimensions: activation, evaluation, and power. We found immediate response patterns, in which the staff member colored her utterances in response to the emotion shown by the student in the immediately previous utterance, and built a predictive model suitable for use in a dialog system to persuasively discuss graduate school with students.

Index Terms: emotional responses, dimensional emotions, user modeling, response strategies, persuasion

1. Aims, Domain, and Corpus

Although spoken dialog systems are becoming widespread, their deployment is today largely limited to domains involving simple information exchange. We would like to develop dialog systems able to support more challenging dialog types, including guidance, decision support, collaborative action, and persuasion. To do so requires new capabilities, including the ability to model the sort of interpersonal interaction that occurs moment-by-moment in human-human dialog. We believe that tracking the user's state, in many dimensions, and displaying understanding of this state immediately is important for efficient and effective dialog. Previous work has shown that this can be done and that users like it [1, 2], but these demonstrations worked in very simple domains, and used hand-coded rules.

We chose to study these phenomena in the domain of persuasion. Persuasion is an interesting domain in that no spoken dialog systems today attempt this, yet spoken dialog is considered to be an especially effective medium for persuasion, often more powerful than print or website appeals, as indicated by the size of the call center workforce involved in sales.

Our specific domain is that of persuading undergraduates that continuing on for a graduate degree is an option worth considering. We worked with the department's graduate coordinator, whose job functions included talking to undergraduates and helping to grow the graduate programs. This staff member was unusually personable and pleasant to talk to; an exemplar for effective dialog behaviors. We brought in 10 students to talk with her, compensating them with credit for one of the assignments in their Introduction to Computer Science class. The students had little knowledge of the nature or value of graduate school and of the application process. The conversations lasted 9–20 minutes. Table 1 shows an excerpt of the corpus (the annotations are explained later).

Preliminary analysis revealed that the dialogs were only mildly persuasive. There were no attempts to get the students to perform any immediate actions, and the dialogs were mostly about providing factual information, although of course the graduate school option was presented in a generally positive way. It was clear that the coordinator was adapting her presentation of the facts to the individual student, selecting content and offering encouragement and guidance based on the student's specific background, status, and concerns. The dialogs seemed appropriate for the students' needs and interests. We know that some of the participants did indeed apply for graduate school a few years later.

Of the various common persuasive strategies [3, 4], the one mostly clearly present was enabling the student to self-persuade. We also saw the use of similarity; as discussed below, the coordinator often mirrored the students' attitude at various points in the dialog.

We set out to reverse-engineer these dialogs. First we identified the common chunks of content. We then created a non-speech version that took checkbox input and generated a customized letter about graduate study. User comments increased our confidence that we had correctly captured the main points of discussion, and also our confidence that presentation of the information by voice, although much slower, would be more effective. We then built a baseline speech version in VoiceXML. This presented content chunks appropriate for the user's answers to simple closed questions. Experiments with 4 users revealed that the system was perceived fairly positively, but that there were several problems, including that the content chunks were too long, which was of course because they were designed that way due to the lack of VoiceXML functions to support smooth and rapid turn-taking.

Two users indicated that the tone of the utterances sounded bored, sad, and without feeling. In contrast, the dialogs in the corpus exhibited clear variation over the utterances of both coordinator and students in emotional coloring and prosodic properties, including pitch, timing, and volume. These did not appear to be accidental or random, rather they seemed to be the primary way that the coordinator showed attention, involvement, and empathy. The dialogs with the VoiceXML systems lacked this, and the overall impression was very different; it was as if the users were browsing a collection of audio clips, rather than having a real interaction. Thus, we developed our topic: modeling the moment-by-moment interaction.

The rest of the paper is structured as follows. After a brief review the literature, we describe our annotation of the corpus, the responsive strategies found, and the learning of a set of rules for predicting what emotional stance the persuasive sys-

Table 1: Annotated excerpt from the persuasive dialog corpus

Line	Transcription	Emotion (Act., Val., Pow.)	Notable Acoustics
GC0	So you're in the 1401 class?	(35, 10, 35)	normal speed, articulating beginnings of words
S1	**Yeah.**	*(10, 5, –5)*	*higher pitch*
GC1	Yeah? How are you liking it so far?	(40, 10, 35)	medium speed, articulating beginnings of words
S2	*Um, it's alright, it's just the labs are kind of difficult sometimes, they can, they give like long stuff.*	*(5, -10, -15)*	*slower speed, falling pitch*
GC2	Mm. Are the TAs helping you?	(20, –10, 10)	lower pitch, slower speed
S3	*Yeah.*	*(5, 5, –15)*	*rising pitch*
GC3	Yeah.	(20, 5, –15)	rising pitch
S4	*They're doing a good job.*	*(10, 0, 5)*	*normal speed, normal pitch*
GC4	Good, that's good, that's good.	(35, 10, 40)	normal pitch, normal speed

tem should take in response to the state of the user as revealed by the tone of his previous utterance. We conclude with discussions of needed improvements and plans for incorporating this into a persuasive system.

2. Related Research

Techniques for inferring users' emotions from their voices have seen remarkable development over the past few years, however, applications of this are still few.

Of particular utility seems to be a dimensional representation of emotion, e.g. [9], using the three dimensions of activation (active/passive), evaluation (positive/negative), and power (dominant/submissive). This enables the representation of more subtleties of emotional state, as reflected in particular by the "emotional coloring" of speech.

In general modeling the user's current state is important for choosing the next system action. Recent work has shown that the prosody of the user's utterance can be a valuable source of information for doing this; for example, it is better for the system to give more explanation when the student's level of uncertainty, as indicated by tone of voice, is higher [5].

Another line of research, with roots in communications and psychology, highlights the importance of various kinds of accommodation between the interlocutors over the course of a dialog. This accommodation, often a convergence, is seen especially in non-verbal behaviors [6, 7]. For example, when meeting a person who displays sadness in their voice, someone wanting to reduce social distance may modify their intonation, speed, and loudness in voice in order to sound empathic.

Gratch and others, focusing on the listening behaviors needed to achieve rapport, have also shown that noticing details of the user's prosody and responding promptly (even before turn-end) can not only improve the perception of the system but also help users talk more [8].

Thus the time is ripe for a study of how to respond to user emotions by means of suitable emotional coloring for the responses.

3. Methods

We analyzed six dialogs from the persuasive dialog corpus.

We separated the speech into utterance units. An utterance unit starts when a speaker begins a turn and ends when either the other speaker interjects or begins a turn. Speech, in times of overlap, is also considered a separate utterance unit. We also separated turns into up to two utterance units when a speaker drastically changed acoustic features in voice.

To determine whether there were perceivable emotions in the corpus, we asked two judges to independently label the utterance units. The corpus was recorded in stereo, one channel per speaker, and the channels were annotated separately. In each dialog, the judges first labeled the coordinator utterance units and then the subject utterance units. The following were given as definitions for each dimension of emotion:

- Activation (also known as activity or arousal) (Passive/Active) If a speaker is active, it sounds like he/she is engaged and shows interest in voice. A passive voice would sound like a lack of engagement or interest. Also, the speaker may sound ready to take action.
- Valence (also known as pleasure or evaluation) (Negative/Positive) This dimension represents the sound of pleasure in the voice. Positive may be shown by sounding upbeat or pleasant, whereas negative may sound down or unpleased.
- Power (also known as dominance or control) (Submissive/Dominant) A dominant sounding voice can sound like the speaker is taking control or is very sure of what he/she is saying. A submissive voice sometimes sounds like there is uncertainty or like he/she is trying to not show too much power in voice.

In this corpus valence typically related to the speaker's attitude towards the entities and topics discussed, for example the teaching assistants, standardized tests, and financial aid. Each dimension was labeled on a continuous scale ranging from -100 to +100 indicating the emotional coloring. For example, a label of -100 for activation meant extremely passive. The judges were asked to listen to each utterance unit at least three times, each time they labeled a single dimension. The inter-judge correlations for each dimension: 0.58, 0.42, and 0.62 for activation, valence and power respectively. The utterances where the ratings disagreed substantially were mostly very short, disfluent, laughter, non-lexical, or corrupted by microphone noise.

Table 1 shows the annotations obtained from our first judge, and includes some non-systematic observations about salient phonetic and prosodic features.

To determine whether the coordinator was reacting to students' emotional state, we grouped one student utterance unit

with one coordinator utterance unit and considered this an adjacency pair. In the normal case, an adjacency pair consisted of an utterance unit by the student and a subsequent response by the coordinator. As a special cases, if both spoke simultaneously, the coordinator's utterance unit was treated as a response to the student's, as it seemed that her adaptation was fast enough for there to be a causal relation even in such cases. In total there were 962 adjacency pairs across the six dialogs.

4. Immediate Response Patterns

We hypothesized that there were "immediate response patterns" determining how the coordinator chose an emotional coloring for her utterance unit in response the emotion expressed by the student in the immediately previous utterance unit. In particular we expected to see evidence of emotional mirroring, as described by Communication Accommodation Theory [6], that is, a matching of the nonverbal features of the student and the coordinator response. We looked for these first by computing correlations across the adjacency pairs; the results are given in Table 2.

Table 2: Correlation coefficients between coordinator emotion dimensions and subject emotion dimensions in adjacency pairs. Results in bold are significant ($p < 0.05$)

		Student		
		Activation	Valence	Power
Coordinator	Activation	**–0.14**	**0.14**	**–0.24**
"	Valence	0.04	**0.34**	-0.05
"	Power	**–0.15**	**0.12**	**–0.31**

In the valence dimension there was clear evidence for mirroring: the correlation coefficient was 0.34. This makes sense: if the student is positive about something the coordinator will tend to take that perspective, and similarly for negative feelings. An example appears in adjacency pair S2-GC2, where the subject speaks slower and with a falling pitch (which sounds negative) and the coordinator (GC2) mirrors his negative voice. Of course this pattern does not mean that the coordinator slavishly mimicked the student's attitudes, however it was common for her to at least acknowledges his feelings before going on. For example, in response to a student who expressed a negative attitude about the financial burdens of graduate school, she first acknowledged that money was a serious concern, in a sombre voice, but in subsequent utterances turned positive as she explained the opportunities for funding.

In the power dimension there was an inverse relationship in power, a –0.31 correlation: if the student sounded dominant, the coordinator generally became more submissive and vice versa. This was probably mostly a reflection of the natural give-and-take of a dialog: when one person is taking the floor, the other person is yielding it. For example, in adjacency pairs GC0-S1 and GC1-S2 the coordinator is clearly leading the conversation and the student following. This pattern also is not invariable; in S3-GC4 it appears that the student's *yeah* is submissive in the sense that he wants to say no more on this topic, but the coordinator thwarts him by also disclaiming any attempt to take the floor, forcing him to make a more explicit statement in S4.

In the activation dimension, the picture is less clear; again there was a negative correlation, but a much weaker one. In fact, the coordinator's activation seems to relate more to the student's power: as the student sounds more dominant, the coordinator becomes more passive (–0.24 correlation).

Of course these simple patterns do not tell the whole story. For example, listening to the dialog in Table 1 we noticed some other things going on in the various emotional dimensions. In GC1, the coordinator starts by showing activeness and dominance, while displaying a slightly positive voice, probably to sound polite and interested, but not overly positive and superficial. In both S1 and S3 the student says only *yeah*, with similar triple values, but since the first responds to a factual question, and the second to a request for an opinion, S3 overall seems significantly less certain. In S4, the subject responds with slightly higher power and more explicit words which seems to enable the coordinator to close out this topic and return to her normal emotional state (active, positive, dominant) in GC5 as preparation for the introduction of a new topic.

5. Building a Predictive Model

Clearly the coordinator is executing some emotionally responsive strategies during these dialogs. While it is interesting to examine such strategies, as above, ultimately our aim is to build a system to determine appropriate responses, and for this we think that machine learning of appropriate rules holds more promise than a labor-intensive study of specific strategies.

Thus we applied machine learning methods in attempts to build a predictor for the coordinator's emotional responses observed in the adjacency pairs. The students three emotion dimensions were taken as attributes and were used to predict the coordinators emotional responses, using only the annotations by the first judge. We used several machine learning algorithms from WEKA [10], with 10-fold cross validation, and measured the correlations between the predictions of the model and the actual values in the corpus. The best performing algorithms were REPTrees and Bagging with REPTrees. Table 3 shows the results.

Table 3: Correlation coefficients between actual dimension value and predicted dimension value using select student dimension levels as attributes, with the highest correlations in bold

Attributes	**Predicted**	**Prediction Correlations**	
Student Dimensions	**Coordinator Dimension**	**REPTree**	**Bagging**
Act, Val, Dom	Act	**0.24**	0.19
Val, Dom	Val	0.28	**0.35**
Act, Val	Dom	**0.34**	0.30

Overall the results are promising: it is possible to predict, to some extent, the emotional coloring to use based only on the emotions expressed in the previous utterance.

Wondering why the results were not higher, we averaged the absolute errors for each dialog, and found that the first dialog (which was collected before the others) had the highest average absolute errors in all dimensions. The student in this first dialog seemed to have a distinct speaking style (West Coast accent and persistent creaky voice); another likely factor was that the persuader was probably still devising her strategies during this first meeting.

Across all speakers, one characteristic of the worst predicted coordinator responses was poor recording quality, when one of the interlocutors was fidgeting or was too far from the microphone or otherwise sounded muffled. In addition, overlapping utterances and short utterances (less than one second) were common in the poorly predicted cases. On the other hand, listening to the best predicted pairs, we found that the utterance units were generally longer, clearer and not overlapping.

Thus, we see that it may indeed be possible to enhance current state-of-the-art dialog systems, not only by focussing on choosing the optimal dialog content, but also by modeling more human-like nonverbal behaviors such as varying prosody to show appropriate emotional responses.

6. Future Work

We would like to examine other factors involved in choosing emotionally-appropriate responses. In this corpus it is almost certainly the case that the emotional responses were not based only on the immediate context. The coordinator may have been adjusting her responses depending on aspects of the students, such as personality, gender, age, and social status (freshman, sophomore, junior, or senior). The coordinator's responses may also have been based on her cumulative interpretation of the user's state. It is also likely that the emotional coloring of her responses was influenced by the student's lexical content.

Regarding utterance boundaries, we would like to explore a finer grained model to complement the utterance-by-utterance response strategies modeled here. Some utterance units had drastic acoustic variation, for example, the coordinator started slow and soft, then immediately changed to fast and loud. We plan to relate emotional interplay using a fixed time periods, instead of by turns, to find more immediate dependencies in the coordinator's response.

We will also test whether the strategies identified here are also found in other domains and dialog types. We will also determine whether the three dimensions are in fact adequate to capture all, or most, of the significant variations in prosody during speech.

Beyond improvement to the model, we plan to work towards the main goal of this research, which is to determine if these strategies are associated with rapport and if it is possible to create a spoken dialog system that can build rapport with users. To do this we will need to train a system to detect emotions automatically from acoustic features of the student's utterances. The output of this can then be used as input to our predictive model to determine the appropriate emotional coloring for the system's next utterance. This can then be used to parameterize a synthesizer, for example MaryTTS, to infuse the lexical content with appropriate emotional coloring. Initially, we plan graft this processing pathway (the bottom path in Figure 1) onto a simple information-giving dialog system; later we will explore deeper integration (the dotted lines in the Figure), in which the detected student emotions may affect the system's information presentation strategy, and conversely the emotional coloring the system needs to achieve may affect the content and lexical form of the utterances. With the integration of these components we will produce a dialog system able to interact with students and persuade them to consider the graduate school option.

7. Acknowledgements

We would like to thank Anais Rivera and Sue Walker for the collection of the persuasive dialog corpus and the members of the Interactive Systems Group for their comments and suggestions. This work is supported in part by the Army ACTEDS program, NSF grant IIS-0415150, and by RDECOM via USC ICT.

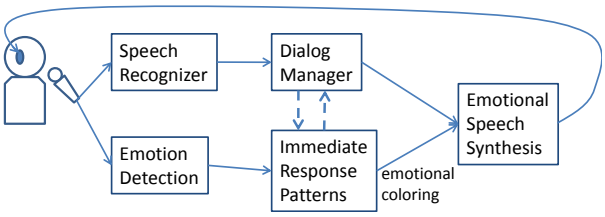

Figure 1: Envisioned spoken dialog system.

8. References

[1] N. Ward and W. Tsukahara, "A study in responsiveness in spoken dialog," *International Journal of Human-Computer Studies*, vol. 59, pp. 603–630, 2003.

[2] N. G. Ward and R. Escalante-Ruiz, "Using subtle prosodic variation to acknowledge the user's current state," in *Interspeech, submitted*, 2009.

[3] J. T. Cacioppo and R. E. Petty, "Language variables, attitudes, and persuasion," in *Attitudes towards language variation*, E. B. Ryan and H. Giles, Eds. Edward Arnold Publishers, 1982, pp. 189–207.

[4] B. Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, 2003.

[5] K. Forbes-Riley and D. Litman, "Investigating Human Tutor Responses to Student Uncertainty for Adaptive System Development," in *Affective Computing and Intelligent Interaction Second International Conference, ACII 2007*. Lisbon, Portugal, September 12-14, 2007: Proceedings. Springer, 2007.

[6] C. Shepard, H. Giles, and B. Le Poire, "Communication accommodation theory," *The new handbook of language and social psychology*, pp. 33–56, 2001.

[7] T. Chartrand and J. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999.

[8] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. van der Werf, and L. Morency, "Virtual Rapport," *Proceedings of the 5th International Conference on Interactive Virtual Agents (IVA)*, 2006.

[9] M. Schröder, *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. Institut für Phonetik, Universität des Saarlandes, 2004.

[10] I. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," *ACM SIGMOD Record*, vol. 31, no. 1, pp. 76–77, 2002.