Achieving Rapport with Turn-by-Turn, User-Responsive Emotional Coloring

Jaime C. Acosta, Nigel G. Ward

Computer Science, University of Texas at El Paso
500 West University Avenue, El Paso, Texas 79968 USA

Abstract

People in dialog use a rich set of nonverbal behaviors, including variations in the prosody of their utterances. Such behaviors, often emotion-related, call for appropriate responses, but today's spoken dialog systems lack the ability to do this. Recent work has shown how to recognize user emotions from prosody and how to express system-side emotions with prosody, but demonstrations of how to combine these functions to improve the user experience have been lacking. Working with a corpus of conversations with students about graduate school, we analyzed the emotional states of the interlocutors, utterance by utterance, using three dimensions: activation, evaluation, and power. We found that the emotional coloring of the speaker's utterance could be largely predicted from the emotion shown by her interlocutor in the immediately previous utterance. This finding enabled us to build Gracie, the first spoken dialog system that recognizes a user's emotional state from his or her speech and gives a response with appropriate emotional coloring. Evaluation with 36 subjects showed that they felt significantly more rapport with Gracie than with either of two controls. This shows that dialog systems can tap into this important level of interpersonal interaction using today's technology.

Key words: affective computing, dimensional emotions, prosody, persuasion, prediction, immediate response patterns, responsivenness, user modeling, social interaction in dialog, interpersonal adaptation

☆We thank Anais Rivera and Sue Walker for the persuasive dialog corpus, Rafael Escalante-Ruiz for the persuasive letter generator, Alejandro Vega, Shreyas Karkhedkar, and Tam Acosta for their help during testing, Ben Walker and Josh McCartney for running subjects, and David Novick, Steven Crites and Anton Batliner for discussion and comments. This work was supported in part by an ACTEDS Scholarship, by NSF Awards IIS-0415150 and IIS-0914868, and by the US Army Research, Development and Engineering Command via a contract to the USC Institute for Creative Technologies.

Email addresses: `jcacosta@miners.utep.edu` (Jaime C. Acosta), `nigelward@acm.org` (Nigel G. Ward)

URL: `http://isg.cs.utep.edu/members/jaime/` (Jaime C. Acosta), `www.nigelward.com` (Nigel G. Ward)

1. Aims

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 of course better speech recognition, speech synthesis and dialog management, but the performance of these core technologies is improving rapidly, and we are now approaching the point where other factors are limiting the further use of dialog systems. Among the new features needed is the ability to engage in the sort of interpersonal interaction that occurs moment-by-moment in human-human dialog. This requires tracking the user's state, in many dimensions, and promptly displaying understanding of this state. We believe that this is important for efficient and effective dialog. Previous work has shown that this can be done and that users like it (Ward and Tsukahara, 2003; Ward and Escalante-Ruiz, 2009), but these demonstrations used hand-coded rules and worked in very simple domains

Building on recent advances in the understanding of emotion in dialog and its prosodic correlates, we have developed a system that brings a new type of responsiveness to human-machine dialog. This paper is structured as follows: after a brief review of the literature, it describes our corpus of persuasive dialogs and its annotation, the responsive strategies found, the system built, the experiments, the results, and some implications.

2. Background

The study of how emotion is expressed by properties of speech has advanced greatly over recent years (Scherer, 2003; Devillers et al., 2005; World-Wide Web Consortium, 2009; Schuller et al., 2009; Batliner et al., 2011; Pittermann et al., 2010), motivated in part by dissatisfaction with the perceived cold, robotic nature of most spoken dialog systems. In particular, it has become possible to infer users' emotions from their voices, thanks to models of acoustic and prosodic correlates of the various emotions, initially for classic emotions such as sadness, anger, and joy in acted speech, and more recently for subtle emotions that are more common in normal spontaneous conversations. Of particular utility for describing these emotions seems to be a dimensional representation (Winton, 1990; Schröder, 2004), using the three dimensions of activation (active/passive), evaluation (positive/negative), and power (dominant/submissive). Dimension-based descriptions cover so much of the prosodic expressiveness in dialog that they also cover states, attitudes and feelings that are not emotions in the strict sense. Thus we will speak of the "emotional coloring" of utterances, rather than their "emotional content," and the word "emotion" will be shorthand for "emotion-related user states" (Batliner et al., 2011).

Despite this improved understanding of emotion in speech, the field has yet to reap the fruits: there have been only a few demonstrations of how to exploit such knowledge in dialog systems (Beale and Creed, 2009; Forbes-Riley and Litman, 2011). One illustration is that of monitoring for user frustration and when detected, apologizing or performing other mitigating actions (Klein et al., 2002). A second illustration is that of monitoring for urgency in the user's voice, words, and turn-taking behavior, and when this is detected skipping the confirmation step to give the desired information immediately (Komatani et al., 2005). A third illustration is in the domain of tutoring dialogs, where Forbes-Riley and Litman (2011) showed that monitoring the user's level of uncertainty, which can be inferred from prosody (Pon-Barry, 2008), and responding to uncertain answers with more explanation was better liked and improved learning.

Forbes-Riley and Litman's work comes close to our goal of moment-by-moment responsiveness, or "dynamic adaptation" as they call it. Here we are interested in going further, in two directions: first, in extracting and using a richer emotion set, with more than one dimension of user state, and second in responding in the matching modality, not with content adjustments but by adding suitable emotional coloring to the system's utterances.

Other inspiring lines of research relate to social interaction (Bevacqua et al., 2008; Saerbeck et al., 2010; Kopp, 2010) and rapport in dialog. Rapport does not necessarily involve deep or lasting feelings between the interlocutors, but includes the sort of positive first impression that may arise from a one-off interaction (Grahe and Bernieri, 1999; Gratch et al., 2006). Responsive listening behaviors are important for the establishment of rapport, and Gratch et al. (2007) have shown that a system which picks up on the user's need for feedback (expressed largely by certain prosodic patterns) and responds promptly (even before turn-end) by nodding can improve rapport. Rapport in turn is implicated in various desireable outcomes, including users' preference, their comfort level with a system and their ability to express things clearly. We would like to go further, by determining not when to respond but how.

In sum, our aim here is to close the circle, to build on the body of basic research on responsiveness and emotion in speech to show how to create dialog systems where these abilities provide value to users.

3. Corpus

We chose to study these phenomena in the domain of persuasion for two reasons. First, persuasion is an interesting domain, as spoken dialog is an especially effective medium for persuasion (Mazzotta et al., 2009), often more powerful than print or the web, as attested by the size of the call center workforce involved in sales. Second, we had in hand a corpus of persuasive dialogs. This we had gathered back in 2005 with the intention of examining turn-taking patterns. However we suspected that they would also be interesting in terms of emotion and rapport.

This corpus consists of 10 dialogs between students and the program coordinator, a department staff member who we had noticed as unusually personable and pleasant to talk to, an exemplar for effective dialog behaviors. Her job functions included giving career advice to undergraduates and helping to grow the graduate program, so it was natural to ask her to talk to undergraduates about graduate school. 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 or of the application process. The conversations lasted 9–20 minutes. Figure 1 shows an excerpt of the corpus (the annotations will be explained later).

Figure 1: Annotated excerpt from the persuasive dialog corpus.

Line	**Transcription**	**Emotion**	**Salient Acoustic Properties**
C0	So you're in the 1401 class?	act: 35, val: 10, pow: 35	normal speed, articulating beginnings of words
S1	*Yeah.*	*act: 10, val: 5, pow: –5*	*higher pitch*
C1	Yeah? How are you liking it so far?	act: 40, val: 10, pow: 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.*	*act: 5, val: –10, pow: –15*	*slower speed, falling pitch*
C2	Mm. Are the TAs helping you?	act: 20, val: –10, pow: 10	lower pitch, slower speed
S3	*Yeah.*	*act: 5, val: 5, pow: –15*	*rising pitch*
C3	Yeah.	act: 20, val: 5, pow: –15	rising pitch
S4	*They're doing a good job.*	*act: 10, val: 0, pow: 5*	*normal speed, normal pitch*
C4	Good, that's good, that's good.	act: 35, val: 10, pow: 40	normal pitch, normal speed

3.1. Preliminary Observations

These dialogs were persuasive in that the coordinator's overall aim seemed to be to persuade the students that continuing on for a graduate degree was something worth considering. However there was no hard sell: 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, and the conversation usually outlined the benefits of graduate school and mentioned specific things the student could think about or do. Of the various common persuasive strategies (Cacioppo and Petty, 1982; Fogg, 2003), the one mostly clearly present was enabling the student to self-persuade: it was clear that the coordinator was adapting her presentation to the individual student, selecting content and offering encouragement and guidance based on the student's specific background, status, and concerns. Similarity was also used: as discussed below, the coordinator often mirrored the students' attitude at various points in the dialog. While we did not directly measure the participants' perceptions of the dialogs nor their effectiveness, we know that some of them did indeed apply for graduate school a few years later, which strengthens our belief that these dialogs had persuasive value.

There are several reasons why dialog may be a good medium for persuasion. Based on the literature, and on comparison with a non-speech approach to delivering information (a system produced that took checkbox input and generated a customized letter about graduate school options), it seems that dialog has the potential for more engagement, more time spent on the topic, more authenticity and believability, and the development of the student's own understanding as a side-effect of articulating his feelings (Holtgraves and Kashima, 2008).

To see how closely we could approach the style and persuasive power of these dialogs with the commercial state of the art, we built a VoiceXML-based system. To do this we identified the common chunks of content across the dialogs, extracted them, and added dialog scaffolding to ask the users simple closed-form questions and then deliver the relevant content chunks. In informal experiments with four users the system was perceived fairly positively, but weaknesses were noted. One was the problem of over-long content chunks, which was due ultimately to the lack of VoiceXML functions to support smooth and rapid turn-taking. This we chose not to address here.

Emotion also was flagged as an issue: two users indicated that the tone of the utterances sounded bored, sad, and without feeling, as can be heard in Audio Figure 1[1]. In contrast, the dialogs in the corpus exhibited clear variation over the utterances of both coordinator and students in emotional coloring and prosodic properties. 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, as illustrated in Audio Figure 2. 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 fixed our topic: modeling the emotional interaction.

3.2. Emotion in the Corpus

In order to analyze this type of interaction, we decided first to annotate the corpus for emotions. Although this was not a difficult decision, it was not the only possible approach. One alternative would have been to focus on domain-related feelings or functions, such as seeking or giving approval. Another alternative would have been to predict the coordinator's prosodic properties directly from those of the context, bypassing the need for emotional mediating variables at all (Ward and Escalante-Ruiz, 2009). However we opted to analyze the dialogs in terms of general emotions: potential advantages include simplification of the model, more intuitions about the regularities found, and leveraging previous work. Even within the emotion-based approach, there are many possible ways to proceed (Calvo and D'Mello, 2010). As our aim in this was not to support conclusions directly from the annotations, we made some rather ad hoc choices, trusting to the ultimate success or failure with the implementation to show whether our analysis was correct.

[1] also available at http://isg.cs.utep.edu/members/jaime/2010specom/

First we attempted to use classical emotions — sadness, happiness, surprise, and so on — but abandoned this due to poor agreement among judges. We then adopted a dimensional approach, using the three generic dimensions. For the annotators we prepared the following colloquial definitions:

Activation (+100=Extremely Active, –100=Extremely Passive) If a speaker is active, it sounds like he/she is engaged and shows interest in his or her voice. A passive voice would sound like a lack of engagement or interest.

Valence (+100=Extremely Positive, –100=Extremely Negative) 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 (+100=Extremely Dominant, –100=Extremely Submissive) 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.

The annotators were asked to assign labels to utterances. Although using utterances as the unit for labeling is not unproblematic, we did this for convenience, taking an utterance to be a segment of speech that starts when a speaker begins a turn and ends when the other speaker either interjects or begins a turn. By this definition, a fragment of speech that overlaps speech by the interlocutor becomes a separate utterance.

As the speakers varied significantly in their vocal and emotional ranges, before working on each track the annotators listened to a random list of utterances of that speaker, to become accustomed to that speaker's style. To avoid the distraction of switching between speakers, in each dialog the annotators labeled all the utterances in one track before going to the other track. Specifically, first they labeled all of the coordinator's utterances then all of the student's utterances. Thus the presentation was partially decontextualized. The annotators were asked to listen and label to each utterance three times, first to assign a value for activation, second for valence, and third for power.

Two annotators were used. They worked independently. Values were directly entered in a spreadsheet, without any special tools. The range was –100 to +100, following Cowie et al. (2001). As a measure of their agreement, we computed inter-annotator correlations for each dimension. For activation this was 0.58, for valence 0.42, and for power 0.62; such levels of agreement are not atypical (Schuller et al., 2009). Being concerned nevertheless that the correlations were not higher, we examined the utterances where the ratings disagreed substantially. It turned out that these were mostly very short, disfluent, laughter, non-lexical, or corrupted by microphone noise; there were almost no large disagreements on normal full turns. We therefore opted not to further analyze or reconcile the labels, as perfecting them was not our aim, and we deeming them good enough to try to use for system development. In fact, we started system development without even waiting for the second annotator to finish, and so references to "annotations" below refer to the first annotator's ratings. Figure 1 shows a dialog fragment with annotations. The whole exchange can be heard as Audio Figure 2, and the individual utterances as Audio Figures 3 through 11.

The annotations revealed substantial emotional variation in each dimension, with standard deviations ranging from 12 to 41 points, both for for the coordinator and for the students, with valence showing the least variation. The activation and power dimensions were similar, correlating at 0.83, with valence being more independent, correlating 0.38 with activation and 0.30 with power.

In this corpus emotions generally did not appear to be deeply felt; rather they appeared to be largely social and to relate fairly directly to the course of the dialog, the interaction with the interlocutor, and the topics being discussed. In particular, activation often seemed to relate to degree of interest in a topic and amount of involvement in the conversation itself; valence to the speaker's attitude towards the entities and topics discussed, for example the teaching assistants, standardized tests, and financial aid; and power to degree of knowledge about a topic, willingness to take the lead in introducing or closing out a topic, and willingness to take or assign the turn.

4. Immediate Response Patterns

It seemed to us that much of the charm and effectiveness of our exemplary speaker came from the way that she was aware of what the students were feeling, of the attitudes and concerns behind their (sometimes laconic) words; and in the way that she showed this awareness in her own words and voice.

As a first step to modeling this, we decided to start simple: to examine whether and how the emotional coloring of her voice in each utterance depended on the emotional coloring of the student's immediately previous utterance. To do this we grouped each coordinator utterance with one student utterance and considered this to be adjacency pair. In the normal case, an adjacency pair consisted of an utterance by the student and a subsequent response by the coordinator. In this framework it wasn't clear what to do with overlaps; we handled such such cases by arbitrarily grouping the coordinator's utterance with the student's simultaneous utterance, and sometimes it appeared that her reactions were swift enough for this to make sense. Six dialogs were processed in this way (with the other four left unexamined, as a resource for future evaluation studies); giving a total of 962 adjacency pairs.

Taking inspiration from Communication Accommodation Theory and other discussions of alignment (Shepard et al., 2001; Suzuki and Katagiri, 2007; Branigan et al., 2010), we looked for evidence of emotional convergence, that is, a matching of the nonverbal features of the student and the coordinator response (Chartrand and Bargh, 1999). While accomodation is typically thought of as operating over the course of a dialog, we looked for immediate effects, by computing correlations between the values of the utterances in each adjacency pair. The results are shown in Table 1.

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 Figure 1, in adjacency pair S2-C2, where the subject speaks slower and with a falling pitch (which sounds negative) and the coordinator (C2) mirrors his negative voice. Of course the

Table 1: Correlation coefficients between coordinator emotion dimensions and subject emotion dimensions in adjacency pairs.

		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

coordinator did not always mimic the student's attitudes, however it was common for her to at least acknowledge 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, a –0.31 correlation, meaning that 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 pairs C0-S1 and C1-S2 the coordinator is clearly leading the conversation and the student following. This pattern also is not invariable; in S3-C3 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 be related more to the student's power: as the student sounds more dominant, the coordinator becomes more passive or withdrawn (–0.24 correlation), perhaps reflecting a strategy for asserting control in a quiet way.

Thus it seems that the dialogs exhibited "immediate response patterns," in which the coordinator responded to the emotions expressed by the student in the immediately previous utterance.

Of course these patterns do not tell the whole story. For example, listening to the dialog in Figure 1 we noticed some other things going on. In C1, the coordinator starts by showing high 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 prosody and thus similarly annotated emotional coloring, but since the first responds to a factual question, and the second to a request for an opinion, in context S3 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 C4 as preparation for the introduction of a new topic. In general, the deployment of emotional coloring is a complex process depending on many additional factors, notably context and dialog strategies (Burgoon et al., 1995; D'Mello et al., 2009).

But immediate response patterns do explain much of the variation so, rather than examine things more deeply, we decided to go ahead and try to put them to use. We chose to use machine learning to quantify these patterns: the students' three emotion dimensions were taken as attributes and from these we sought to derive predictions of the coordinator's emotional coloring in the next utterance. We tried several machine learning algorithms from WEKA (Witten and Frank, 2005), including MultilayerPerceptron, Support Vector Machine, Linear Regression, and the tree-based models M5PTree and REPTree. Performance was measured using ten-fold cross validation: the data was split into ten pieces, nine of the ten pieces were used for training while the remaining piece was used to evaluate the performance of the trained model, and then the split was rotated to predict each tenth in turn. All of the speaker data was mixed, and thus the test data was never from a completely unseen speaker. The performance measure was the correlation between the predictions of the model and the actual values in the corpus. The best performing algorithms were REPTree and Bagging with REPTree (Witten and Frank, 2002).

Table 2: Correlations by dimension between actual coordinator emotion and predicted emotion, predicted from labeled emotional state in the student's previous utterance, with the highest correlations in bold.

Predicted	Student	Prediction Quality	
Coordinator	Dimensions used	(Correlation Coefficient)	
Dimension	in the Predictor	REPTree	Bagging
Act	Act, Val, Dom	**0.24**	0.19
Val	Val, Dom	0.28	**0.35**
Dom	Act, Val	**0.34**	0.30

Table 2 shows the results. They confirm that it is possible to predict, to some extent, the emotional coloring to use based only on the emotions expressed in the previous utterance.

Curious as to why the correlations were not higher, we examined the utterances whose colorings were predicted poorly. First, we looked across the dialogs, averaging the absolute errors for all the utterances in each. For each dimension the first dialog was the worst. Perhaps for this first dialog the coordinator was behaving differently, possibly being not yet comfortable with the situation or recording setup. Second, we looked at individual utterances. Across all speakers, one characteristic of the worst predicted coordinator responses was poor recording quality, either for the predicted utterance or its predecessor, for example 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, of less than one second, were common in the poorly predicted cases. On the other hand, listening to the best predicted pairs, the utterances were generally longer, clearer and not overlapping.

We therefore decided that the immediate response patterns found were consistent enough to use in a system, and likely important enough to make a difference to users' impressions.

5. System Description

To sum up the findings so far, in our corpus the participants vary their emotional coloring utterance-by-utterance, these variations depend in part on the emotions just expressed by the interlocutor, and these response patterns can be quantified fairly well. Thus we set out to build a system capable of modeling this behavior. This section describes this system, Gracie, the GRAduate Coordinator with Immediate-response Emotions.

Gracie was built using existing open source components as much as possible. Figure 2 shows the overall system architecture. This section begins by describing the heart of the system, the immediate response rules, and works outwards towards the periphery.

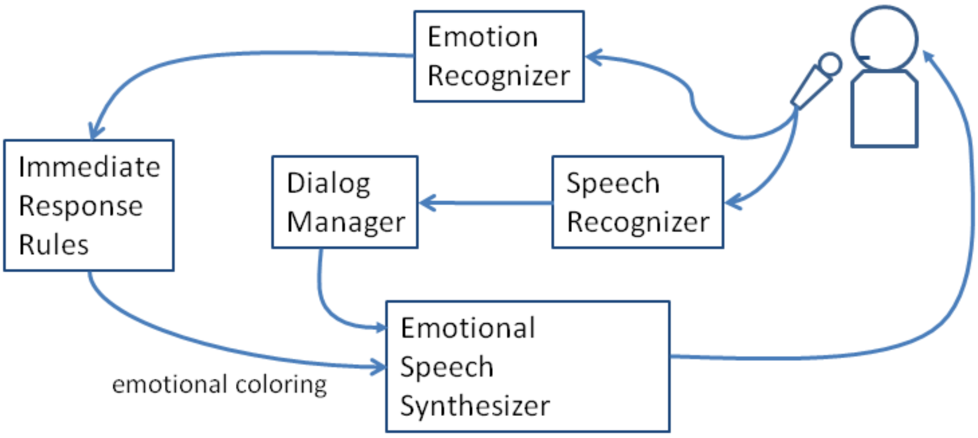

Figure 2: Gracie System Architecture

5.1. Immediate Response Rules

Initially we built the response rules by simply transplanting the equations from WEKA to Gracie. However in pilot runs the resulting system's behavior was sometimes erratic, characterized as "bi-polar" by one tester. We attributed this problem to discontinuities in the REPTree-based response function, where a slight change in the user's emotion could cause a drastic change in Gracie's response. We had originally selected REPTree because it performed best on the corpus, but we decided that this was unacceptable for a live system. We resolved this issue by training a new version using Linear Regression, again using WEKA. The new rules are seen in Figure 3. They performed slightly worse in terms of match to the labels in the corpus (with correlation coefficients of 0.15, 0.32 and 0.30 for activation, valence and power respectively). However exact match to the annotator's labels is probably not important, as these were not evenly distributed but favored certain values, for example multiples of 10, and the Reptree algorithm may have picked up on such unintended patterns and modeled them faithfully at the expense of overall continuity.

```
activation response value = 0.20 * SubjVal - 0.16 * SubjPow + 6.03
valence response value = 0.52 * SubjVal + 1.22
power response value = -0.29 * SubjPow - 4.95
```

Figure 3: Functions used to calculate the system's emotional coloring from the student's emotion in the immediately preceding utterance.

5.2. Dialog Manager and Speech Recognizer

The dialog manager is responsible for the content of the system's utterances. Analysis of the corpus led to the identification of 64 topics, grouped into 19 sections (Acosta, 2009). In the dialog manager each topic consists of a system utterance, a list of words or phrases to recognize, and an action. The system utterances were based on coordinator utterances in the corpus, selected and sometimes adapted to match the desired persona (Acosta, 2009) and the limitations of the restricted turn-taking abilities we could provide. For speech recognition we used PocketSphinx (Huggins-Daines et al., 2006), because it is open source, fast and quite accurate when the set of recognition possibilities is small, as here. Actions specify what the system will say next and indirectly affect what it may say in the future, by enabling or disabling entry into various sections of content based on the user's responses.

As suggested earlier, the emotional responses of the system do not depend on the information in the dialog manager. The reverse is also true: the dialog manager selects the content at each point without considering the user's emotional state nor the planned emotional coloring. This is to keep the system simple. As inconsistency between emotion and content is undesireable (Berry et al., 2005), we avoided this by the simple expedient of choosing system utterances that would be compatible with any emotional coloring.

5.3. Emotion Recognizer

In order for the Immediate Response Rule module to function, it needs information on what the user's previous emotion was. This is provided by the Emotion Recognizer, which infers values for the three dimensions from the user's prosody.

While recent research has begun to identify the features of the voice that are most effective for emotion recognition (Schuller et al., 2009; Batliner et al., 2011), for convenience we used just the 37 features that our software infrastructure happened to compute. These features came from studies of turn-taking in various languages, but, as we know that the prosodic features used in emotion are not disjoint from those used in turn-taking (Ward and Bayyari, 2010), we thought that they would also likely be useable here. These included measures of energy and pitch, various functionals derived from these (slope, average, maximum), and a proxy for speaking rate (Acosta, 2009). Each feature was computed every 10 milliseconds but the system used the averages over each entire utterance. Per-speaker normalization was not done.

Again we used machine learning to find a predictor. We chose to use linear regression because the result of training, a set of linear equations, would be continous and easy to transplant to a real-time component for Gracie. Specifically we adopted M5P (Frank et al., 1998) because it performed best.

Table 3: Correlation coefficients between the output of the Emotion Recognizer and the annotations.

Emotion Dimension	Recognizer Quality (Correlation Coefficient)
Activation	0.73
Valence	0.44
Power	0.79

Table 3 shows the correlations between the predicted values (from the trained model) and the actual values (from the annotator) for each dimension of emotion. Correlation coefficients for activation and power are high, but moderate for valence. In other corpora valence has also been the hardest dimension to predict from prosody (Schuller et al., 2009), perhaps because it depends more on lexical information.

5.4. Speech Synthesizer

To voice the system's output, we needed a synthesizer able to express emotion. MaryTTS with Emospeak (Schröder and Trouvain, 2003) provided the necessary functionality. This module takes as input a sequence of words and a three-dimensional emotion triple and outputs an emotionally colored utterance. Although based on emotion-prosody mappings found for German, its manipulations sounded fine to us for English.

While the range of inputs to Emospeak, –100 to +100, appeared to match the range in our annotations, we discovered that the ranges were not aligned: in a pilot study Gracie often produced output at a speaking rate so fast as to be unintelligible, presumably as a way of expressing extreme joy. We therefore had to scale things to fall within the normal range of emotion expressed in a quiet conversation. We did this by normalizing the labels of the coordinator's emotions to have a mean of 0 and a standard deviation of 10 on each dimension. After this all values were in the range of –30 to +30. We then retrained the immediate response patterns and as a result all predicted emotional coloring values also fell in this range, and informal tests with laboratory members indicated that Gracie's outputs were now in an appropriate range.

5.5. System Integration and Dataflow

Most of the components being in C or C++, we integrated them using Eclipse on the CentOS distribution of Linux. The exception was MaryTTS, which however comes with a socket interface, so we created a C++ client and interfaced to it that way.

Gracie uses traditional, rigid turn-taking, both for convenience of implementation and in order to prevent the occurrence of overlaps, since those would probably not be handled well by our response patterns (Section 4). Specifically, the system produces a prompt, waits for the user to respond, and then waits for at least one second of trailing silence before beginning processing. At this point the utterance is passed to both the dialog manager and to the emotion recognizer. The values output by the Immediate Response Rules then determine the emotional coloring for the next utterance, up to the point when a new set of values is determined after the next user input.

6. Experiment Design

This section describes how we tested our hypothesis, that a spoken dialog system with emotional responsiveness would be better at gaining rapport than a spoken dialog system without it.

6.1. Conditions and Measures

Needing to compare Gracie to systems without emotional responsiveness, we used two controls. The first was Neutral, a version of Gracie with no emotional variation at all: all system utterances were synthesized with values of 0 on all dimensions. A second control was included because users tend to like variation in speech (Ward and Tsukahara, 2003), so we needed to determine whether the benefit was truly coming from the proper selection of emotional coloring, rather than from the mere existence of emotional variation. The second control was Non-Contingent, a version with the same amount of variation (on average) but where this variation was not based on the user's state. Following Gratch et al. (2007), the system determined the sequence of emotional colorings in the non-contingent case by reading them from a file, where the file contained the sequence of triples calculated by Gracie during the interaction with the *previous* subject. Thus the expressions in the non-contingent case do represent a possible and natural sequence of emotional colorings, but just not necessarily one that matches up with the sequence of states of the current subject.

The key judgment obtained from the subjects was their degree of agreement with the statement "I felt I had a connection with the coordinator," on a seven point scale. We took this measure of emotional rapport from Gratch et al. (2007). We also elicited other judgments, including a measure of cognitive rapport and eight other attributes relevant to persuasion and to dialog systems in general, as seen in Table 4.

6.2. System Modifications

To prevent the results being affected by the inevitable occasional misrecognition, for the experiments we turned off the speech recognizer, and compensated by selecting from the system's topics only content for the system-side utterances that would be appropriate no matter what the user said. An example is seen in Figure 4. Thus each dialog progresses in the same way regardless of the words the subject says: the only variation is in the emotional coloring.

Figure 4: One of the fixed content sequences used for the experiment.

Hi, I'm Gracie, tell me your name.

. . .

Nice to meet you.

I'm really glad you called. Now I can tell you about graduate school. There are a couple of things before you get into graduate school. You have to take an exam. The exam is not hard. It is about the same as a high school exam. There is a reading, math, and writing part. It is worth getting into graduate school. Tell me what you think of exams.

. . .

Interesting answer.

Each teacher does research. You can ask each one about their work. Graduate school is about expanding the field of Computer Science. Tell me about your favorite computer science class.

. . .

Good.

Tell me how your teachers and graders help you do better in your classes.

. . .

Oh I see. That is very interesting.

If you want to teach you will need a graduate degree. Usually a University requires a Doctoral degree to teach. Tell me your plans for your future.

. . .

Alright, that's good.

It takes longer to get a graduate degree, but you will enjoy it. You meet great people and do fun research.

. . .

Thanks for listening.

In order to enable subjects to interact with the different versions of the system without experiencing repetition, we needed different content sequences. We devised three, including that seen in Figure 4. All content sequences were similar in length and style (Acosta, 2009). Topics included the application process and the factors involved (grades, statement of purpose), career goals, employment opportunities, pay advantages, funding opportunities for graduate students, reasons to like research, and other reasons to go to graduate school. The content sequences were kept brief, 7 turns each, in order to reduce subject fatigue.

To avoid problems with users overlapping the system's utterances or failing to respond, the system provided guidance by displaying the words "Please Speak" on the screen when it needed input. To avoid problems with users failing to understand what the system was saying, which was an issue when the speaking rate was high, the sentences that the system was saying were simultaneously displayed on the screen.

Thus the experimental situation was unnatural in several respects. We could have done things differently to provide a better user experience, but instead we just made the minimal changes to the fully automatic system. From the user perspective, the result was less like a dialog than we would have liked, and some subjects behaved somewhat oddly, talking to the system in ways that they would never speak to a human. However all did succeed in interacting with the system, and almost all did reveal some degree of emotion in their voice. After all, with prompts like "tell me about your work experience," most subjects naturally revealed at least some feelings.

6.3. Subjects and Procedure

A total of 36 subjects participated in the study, 23 male, 13 female, all students. Twelve were from the Fall 2009 Introduction to Computer Science class, and the remainder, recruited by approaching students in the Computer Science open lab or in the Student Union, were given $10.00 for their participation. None of the subjects had interacted with any version of Gracie before the experiment. As the coordinator had resigned in 2006 and had not been replaced, it is unlikely that any had interacted with her, or indeed had had a serious discussion about graduate school with any of the staff. All but three subjects were bilingual or otherwise reported speaking more than one language.

After completing the consent form, subjects were told that the experiment was a study in communicative technology that would investigate the effectiveness of a dialog system at informing students about the graduate school option. Subjects were told that the task was simply to "interact with Gracie" and give us their opinions. We chose not to measure nor control their expectations further, although we know that expectations affect performance, because the balanced design meant that such effects would not affect the conclusions. We also told the subjects that the system was not recognizing their words at all. After a brief explanation they were given a thirty second demonstration interaction.

Next, the subject interacted with the three systems, filling out a questionnaire and giving comments after each one. In order to focus their attention on the emotional coloring, subjects were asked to base their answers strictly on Gracie's voice, not on the content presented, although the comments showed that they did not always succeed in doing so. As noted above, there were 3 different conditions (rule-based, non-contingent, neutral), and 3 different content sequences. The conditions were presented in all 6 possible orders, and the same for the content sequences. Thus there were a total of 36 orderings of configurations and content sequences. Each subject experienced one of these orderings, thus the ordering of conditions was balanced and so was the match-up between conditions and content sequences.

After finishing the three interactions, subjects were asked to state their preference for each system, from best to worst, asked for additional comments, and then debriefed.

7. Results

Table 4 shows the results for the rating questions (with signficance computed using a paired samples t-test) and Table 5 shows the preference results.

On emotional rapport, question 1, the rule-based version of Gracie was rated higher than either control (in fact higher than non-contingent at p< 0.01) and so our main hypothesis was supported. The effect size was 0.60 standard deviations (computed using the pooled standard deviation of all responses to this question). For all ten of the rating questions, the mean rating for the rule-based system was higher than for either control, although the differences were not always significant.

In the preferences stated after using all three systems, 20 out of 36 subjects preferred the rule-based system over both controls. Comparing with the controls pairwise, 26/36 prefered the rule-based system to the neutral ($p < 0.01$, sign test) and 23/36 prefered the rule-based system to the non-contingent, although this is not significant ($p = 0.07$).

Comments from those who prefered the rule-based system illustrate the type of rapport they felt, for example "[the rule-based system] was more like a conversation rather than just listening," "it seemed [like] talking to a real person," "[the rule-based version] seemed easiest to connect to and stay engaged," and "I was very comfortable with the [rule-based] system." Comments about the non-contingent version were generally less positive. For example one subject commented that "it was very unreal like she made her sentences sound really happy." An example of the importance of adapting to user emotions can be seen by the comments of two consecutive subjects. The first subject praised the rule-based system saying, "the system was able to adapt to the speed at which it was responded to." The following subject said, "it was boring" when the same sequence of emotional colorings were presented non-contingently. Comments about the neutral system were largely negative, for example "[the neutral version] was very monotonous."

Thus, overall, the results indicate that it is important to not only add emotion to spoken dialog systems, but to also adapt the emotions to user states.

However not all subjects prefered the rule-based system. There are several possible reasons for this. One factor was that the lack of variation in prosody meant that the neutral version was most reliably slow, clear and intelligible, which may be why 7 of the 36 liked it best; one stated "[the neutral version] was easier to understand." Another factor may have been personality differences, as suggested in previous work (Ward and Tsukahara, 2003; Cassell and Bickmore, 2003): as our model speaker seemed extroverted, the response patterns based on her behavior may have been less pleasing to subjects more comfortable with a more unvarying conversational style. A final likely factor was the prosodic behavior

Table 4: Subjects' ratings of the three versions of Gracie. In each cell the upper value, in bold, is the mean, and the lower value, in italics, is the standard deviation.

(*) higher than neutral, $p < 0.05$

(**) higher than neutral, $p < 0.02$

(+) higher than non-contingent, $p < 0.05$

(++) higher than non-contingent, $p < 0.02$

Question	**Rule-Based**	**Non-Contingent**	**Neutral**
1 - Emotional Rapport	**++**4.61**	**3.67**	**3.78**
(I felt I had a connection with the coordinator.)	*1.38*	*1.71*	*1.62*
2 - Cognitive Rapport	**+**4.72**	**3.94**	**3.75**
(I think the coordinator and I understood each other.)	*1.61*	*1.45*	*1.92*
3 - Helpful	**4.58**	**4.33**	**4.11**
(The coordinator seemed willing to help.)	*1.75*	*1.67*	*1.75*
4 - Trustworthy	**4.28**	**4.22**	**3.78**
(The coordinator seemed trustworthy.)	*1.65*	*1.87*	*1.77*
5 - Likeable	****4.72**	**4.19**	**3.56**
(The coordinator seemed likeable.)	*1.70*	*1.79*	*1.93*
6 - Natural	**3.94**	**3.58**	**3.19**
(My conversation with the coordinator seemed natural.)	*1.96*	*1.83*	*1.83*
7 - Enjoyable	****4.69**	**3.89**	**3.47**
(I enjoyed the interaction with the coordinator.)	*1.88*	*1.83*	*1.95*
8 - Human-like	****3.94**	**3.31**	**2.78**
(The coordinator was human-like.)	*1.82*	*1.88*	*1.88*
9 - Persuasive	****4.17**	***3.67**	**2.97**
(The coordinator was persuasive.)	*1.89*	*1.82*	*1.70*
10 - Recommendable	**4.28**	**3.89**	**3.64**
(I would recommend the coordinator to others.)	*1.70*	*1.98*	*1.88*

Table 5: Subjects' Preferences for the Three Versions

	Best	Second	Worst
Rule-Based	20	9	7
Non-Contingent	9	12	15
Neutral	7	15	14

of the participants. Of the 19 subjects who listed English as the first of their languages, 13 preferred the rule-based system, however the pattern was different for the others. In particular, of the 9 who preferred the non-contingent version, 8 listed some other language first. This could be because the non-native speakers often seemed to display less prosodic variation, and in such cases Gracie would in turn show less variation. Thus such subjects could actually receive more variation in the non-contingent condition than in the rule-based condition; as suggested by the comments of one non-English-first subject who felt the rule-based system "spoke at strange rate" whereas the non-contingent system "seemed to have more of a personality."

8. Conclusions and Directions

Gracie is the first demonstration of a spoken dialog system that does turn-by-turn tailoring of the emotional coloring of utterances in response to the user's inferred emotional state. The experimental results confirm the hypothesis, that a spoken dialog system with this sort of emotional intelligence is better at gaining rapport with users.

On the one hand, this success came despite many missteps and poor decisions. In many ways the techniques we used were inferior to best-of-class solutions. And yet it worked. Brave and Nass (2007), in their survey of emotion in human-computer interfaces, include in their list of open question "how accurate must emotion recognition be to be useful as an interface technique?" Our result suggests that the answer is "not very." We believe that this is because emotion, at least at this level, is not just a fancy extra, not a mere incidental property of dialog, but part of something fundamental about human interaction (Cappella, 1991), and so even approximate modeling can bring significant benefits. The robust effect size obtained further suggests the power of doing so.

On the other hand, the result is not just luck: there were some things that we think were done right. We feel that the success is attributable, first of all, to the fact that previous research provided almost all the pieces needed. Another key to success was the use of a continuous, rather than categorical, representation of emotion. Another was our choice of time-scale: the immediate response patterns we found are both simple and powerful, whereas emotional interplay at longer time scales is probably more complex and less unambiguously valuable. Finally, we feel that we succeeded because we were willing to simplify and to use a good-enough model rather than trying to be comprehensive or highly accurate.

This is not to say that there is no need for further basic research. Far from it. Our experience in this work confirms the need for a better understanding of, among other things, the dependence of emotion on context (Lee et al., 2009), the relation of emotion to dialog and persuasive strategies, the joint role of prosody and lexical content in conveying emotion, and individual personality differences in emotional feelings and their expression.

In addition to these well-known issues, a newer one which seems critical is that of the time course of emotional responsiveness. Contrary to our working assumption, emotion can vary within an utterance, sometimes drastically (Becker et al., 2004; Batliner et al., 2009). This was a problem both for our annotations and in the responses, where sometimes the system's voice would start out in a highly positive tone, and continue in that exact same tone until the next user response, where it probably would have been better to fade out the positivity after a few seconds. It would be interesting to go beyond utterance-by-utterance modeling to do a serious examination of the time constants of emotional responsiveness in dialog.

Finally, we are interested in examining how these techniques can be used in other domains. As the emotion dimensions used are not specific to persuasion but generic, we are optimistic that in other domains also similar immediate response patterns may improve the user experience, for example in systems for tasks such as obtaining information, ordering services or making travel arrangements. If immediate response patterns in these domains are, as in the persuasion domain, partly independent of the content conveyed, such benefits could be obtained by simply augmenting an existing system with such patterns, with no need to redesign the dialog flow or dialog manager. As the benefits were not just in rapport, but also in aspects which every dialog system needs, the practical utility of responsive emotional coloring could be high.

References

Acosta, J. C., December 2009. Using emotion to gain rapport in a spoken dialog system. Ph.D. thesis, University of Texas at El Paso.

Batliner, A., Seppi, D., Steidl, S., Schuller, B., 2009. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. In: Advances in Human-Computer Interaction. Hindawi.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N., 2011. Whodunnit: Searching for the most important feature types signalling emotion-related user states in speech. Computer Speech and Language 25, 4–28.

Beale, R., Creed, C., 2009. Affective interaction: How emotional agents affect users. International Journal of Human-Computer Studies 67, 755–776.

Becker, C., Kopp, S., Wachsmuth, I., 2004. Simulating the emotion dynamics of a multimodal conversation agent. In: Andre, E., et al. (Eds.), Affective Dialogue Systems. Springer, pp. 154–165.

Berry, D. C., Butler, L. T., de Rosis, F., 2005. Evaluating a realistic agent in an advice-giving task. International Journal of Human-Computer Studies 63, 304–327.

Bevacqua, E., Mancini, M., Pelachaud, C., 2008. A listening agent exhibiting variable behaviour. In: Proceedings of The 8th International Conference on the Intelligent Virtual Agents.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., 2010. Linguistic alignment between people and computers. Journal of Pragmatics 42, 2355–2368.

Brave, S., Nass, C., 2007. Emotion in human-computer interaction. In: Sears, A., Jacko, J. (Eds.), The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, 2nd Edition. Lawrence Erlbaum, pp. 77–92.

Burgoon, J. K., Stern, L. A., Dillman, L., 1995. Interpersonal Adaptation: Dyadic Interaction Patterns. Cambridge University Press.

Cacioppo, J. T., Petty, R. E., 1982. Language variables, attitudes, and persuasion. In: Ryan, E. B., Giles, H. (Eds.), Attitudes towards language variation. Edward Arnold Publishers, pp. 189–207.

Calvo, R. A., D'Mello, S., 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing 1, 18–37.

Cappella, J. N., 1991. The biological origins of automated patterns of human interaction. Communication Theory 1, 4–35.

Cassell, J., Bickmore, T., 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. User Modeling and User-Adapted Interaction 13, 89–132.

Chartrand, T. L., Bargh, J. A., 1999. The Chameleon Effect: The Perception-Behavior Link and Social Interaction. Journal of Personality and Social Psychology 76, 893–910.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollais, S., Fellenz, W., Taylor, J. G., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18, 32–80.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

D'Mello, S., Craig, S., Fike, K., Graesser, A., 2009. Responding to learners' cognitive-affective sta;tes with supportive and shakeup strategies. In: Proceedings 13th International Conference on Human-Computer Interaction, LNCS volume 5612. Springer, pp. 595–604.

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

Forbes-Riley, K., Litman, D., 2011. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. Computer Speech and Language 25, 105–126.

Frank, E., Wang, Y., Inglis, S., Holmes, G., Witten, I., 1998. Using model trees for classification. Machine Learning 32 (1), 63–76.

Grahe, J. E., Bernieri, F. J., 1999. The importance of nonverbal cues in judging rapport. Journal of Nonverbal Behavior 23, 253–269.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., Morency, L.-P., 2006. Virtual rapport. In: 6th International Conference on Intelligent Virtual Agents. pp. 14–27.

Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R., 2007. Creating rapport with virtual agents. In: Intelligent Virtual Agents 2007, LNAI 4722. Springer, pp. 125–138.

Holtgraves, T. M., Kashima, Y., 2008. Language, meaning, and social cognition. Personality and Social Psychology Review.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., Rudnicky, A. I., 2006. Pocket-Sphinx: A free, real-time continuous speech recognition system for hand-held devices. In: IEEE International Conference on Acoustics, Speech and Signal Processing.

Klein, J., Moon, Y., Picard, R. W., 2002. This computer responds to user frustration: Theory, design, and results. Interacting with Computers 14 (2), 119–140.

Komatani, K., Ueno, S., Kawahara, T., Okuno, H. G., 2005. User modeling in spoken dialogue systems to generate flexible guidance. User Modeling and User-Adapted Interaction 15, 169–183.

Kopp, S., 2010. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Communication 52, 587–597.

Lee, C. C., Mower, E., Busso, C., Lee, S., Narayanan, S., September 2009. Emotion recognition using a hierarchical binary decision tree approach. In: Interspeech 2009. Brighton, UK, pp. 320–323.

Mazzotta, I., Novielli, N., Carolis, B. D., 2009. Are ECAs more persuasive than textual messages? In: Ruttkay, Z., et al. (Eds.), Intelligent Virtual Agents. Springer, pp. 527–528.

Pittermann, J., Pittermann, A., Minker, W., 2010. Emotion recognition and adaptation in spoken dialog systems. International Journal of Speech Technology 13, 49–60.

Pon-Barry, H., 2008. Prosodic manifestations of confidence and uncertainty in spoken language. In: Interspeech. pp. 74–77.

Saerbeck, M., Schut, T., Bartneck, C., Janse, M. D., 2010. Expressive robots in education. In: CHI. pp. 1613–1622.

Scherer, K. R., 2003. Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256.

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

Schröder, M., Trouvain, J., 2003. The German text-to-speech synthesis system Mary: A tool for research, development and teaching. International Journal of Speech Technology 6 (4), 365–377.

Schuller, B., Vlasenko, F., Eyben, B., Rigoll, G., Wendemuth, A., 2009. Acoustic emotion recognition: A benchmark comparison of performances. In: Proceedings Automatic Speech Recognition and Understanding Workshop. IEEE, pp. 552–557.

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

Suzuki, N., Katagiri, Y., 2007. Prosodic alignment in human-computer interaction. Connection Science 19, 131–141.

Ward, N., Tsukahara, W., 2003. A study in responsiveness in spoken dialog. International Journal of Human-Computer Studies 59, 603–630.

Ward, N. G., Bayyari, Y. A., 2010. American and Arab perceptions of an Arabic turn-taking cue. Journal of Cross-Cultural Psychology 41, 270–275.

Ward, N. G., Escalante-Ruiz, R., 2009. Using subtle prosodic variation to acknowledge the user's current state. In: Interspeech. pp. 2431–2434.

Winton, W. M., 1990. Language and emotion. In: Giles, H., Robinson, W. P. (Eds.), Handbook of Language and Social Psychology. John Wiley & Sons, pp. 33–49.

Witten, I. H., Frank, E., 2002. Data mining: practical machine learning tools and techniques with Java implementations. ACM SIGMOD Record 31 (1), 76–77.

Witten, I. H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann.

World-Wide Web Consortium, 2009. Emotion markup language (EmotionML) 1.0, W3C working draft, marc Schröeder, ed.