Persuasion, Dialog, Emotion and Prosody

Jaime C. Acosta
University of Texas at El Paso
500 W. University
El Paso, Texas 79968
jaimeacosta@acm.org

ABSTRACT

In this paper, I describe open research questions associated with building a persuasive dialog system that can gain rapport with users by recognizing and synthesizing appropriate three dimensional emotions (activation, valence, and power). I also discuss why such an emotionally intelligent system is well suited for mobile devices, specifically those with voice input.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors

General Terms

Human Factors

Keywords

mobile persuasion, persuasive technology, rapport, emotion, emotion dimensions, immediate response patterns, emotion recognition

1. INTRODUCTION

Since the early 90s affective computing has been used in many domains. For example, there have been systems that utilize emotion for customer support[3], providing tutoring [5], and persuasion[2, 7, 6].

While desktop PCs have taken advantage of emotional computing, research in mobile systems that use emotion is still lacking. Speech processing tools, for example speech recognizers, end of utterance detectors, and emotion recognizers, are becoming more accurate and their availability is reaching the mobile community. Embedded processors are becoming more advanced and are able to run applications that were before strictly for use on high end desktop PCs. The next step in mobile HCI is to focus on the application of better serving user needs based on emotion indicators such as prosody in voice.

One area of research that has recently received much attention is mobile persuasion as seen in the rise of venues such as *MobilePersuasion* since 2007. Fogg [4] mentions that the time and place that information is given is important for persuasion. Mobile devices are intrinsically suited for these types of persuasion because, unlike desktop computers, people have them wherever they go.

The rest of the paper is organized as follows. First I discuss my claims regarding the relationship between persuasion, rapport, and emotion. Next, I describe the development procedures of a persuasive dialog system and how such a system could be implemented for use in a mobile device.

2. CLAIMS

2.1 Emotional sensitivity is needed to gain rapport in mobile applications

When people interact with each other, they share a variety of nonverbal behaviors. Sometimes people that do not react to a speaker's emotion can be seen as careless and even negative. For this reason, artificial agents must be able to determine user emotional state and react appropriately; some research is working on this. Since mobile devices are small, they are limited by their input methods. Voice is a common choice for these devices because it makes it easy for humans to use, without having to dedicate their hands, eyes, and others. For this reason, emotion sensitivity through voice is key for mobile applications. For example, in customer support applications [3], when a user sounds frustrated, the system will transfer the caller to a human.

Regarding gaining rapport, some may think that rapport is gained automatically after knowing someone for a long time. If this was the case with computers, people that use, for example, answering machines for long periods of time, would feel a sense of rapport with the machine. This is not the case. This means that something is missing; I think it is the ability for the machine to detect emotion and adapt acoordingly. In this research I look to gain instant rapport, which means that an artificial agent is able to gain rapport with a user within a short period of time, in my case, within one interaction. Nonverbal behaviors that signal emotions may be key. Spoken dialog systems and artificial agents in general lack emotional intelligence and this makes their artificiality easy to spot.

2.2 Emotional responses are needed to gain rapport in mobile systems

In addition to detecting emotional states of users, it is also important to respond appropriately. If we are to create systems that exhibit more human-like interaction patterns, we must also build systems that respond to users with appropriate actions, dependent on their emotional state. Communication Accomodation Theory [9] says that people change

Copyright is held by the author/owner(s).
MobileHCI'09 September 15-18, 2009, Bonn, Germany
ACM 978-1-60558-281-8.

their nonverbal behaviors in order to reduce social distance. If there these nonverbal behaviors are absent, it could be seen as negative. This may be why spoken dialog systems are seen by many, negatively. They seem without emotion, uncaring for the user's feelings. When dealing especially with persuasion, rapport, gained through emotion, is a key element because it creates a closer relationship with users.

I argue that implementing a system that only adapts information based on user emotion is not suitable. Taking the example of the customer support systems, if the system would simply transfer a frustrated customer to a human person, this may not be enough. It would seem as if the customer was being put aside, and that only now that they are showing frustration, they will be taken seriously. It would be more effective if the spoken dialog system showed emotion throughout and then, when the user seemed frustrated, first apologized in a sympathetic voice, then transferred the user to a human.

2.3 Speech and Dialog is good for persuasion

When people engage in persuasive dialog, their speech is composed of arguments and counter-arguments. This is not possible in one-sided persuasion, such as billboards or even commercials. When persuasion occurs in dialog, people learn about the other person and tailor their arguments. This is probably why we still rely on salesmen, not just advertisements, to sell. We hear arguments tailored to us and colored with emotional prosody.

Although advertisements are good methods of persuasion, dialog may be better, even if the dialog is with a machine. With one-way arguments, people may have preconceived knowledge that may or may not be true. This cannot be dealt with in a one-way argument. Advertisements are effective, but they may mainly serve as a starting point, maybe to catch the interest of the person. Most people find it difficult to, for example, lose weight or commit to an exercise routine, unless they have someone (with whom they probably have rapport) that can encourage them to continuously practice their routine.

2.4 Three dimensions of emotion suffice for most dialogs

Using the three dimensional approach provides us with an easy way to characterize prosodic patterns. The three dimensions and their acoustic correlates, described in [8], are defined as follows: activation is activity in voice or sounding ready to take action. Mean pitch, intensity and speaking rate and others are mostly correlated with activation. Valence is either positive or negative, where positive valence is correlated mostly with lower pitch, large pitch range and others. Power is defined as sounding dominant or submissive. The acoustic correlations of power are similar to activation, except that lower pitch represents dominance (whereas higher pitch represents activity). Using the three dimensions, we can build recognizers and synthesizers and not have to account for every aspect of the spoken signal; instead they can focus on these limited correlates in the acoustic signal. This makes it possible to run emotionally sensitive applications on real-time systems and limited resource systems, such as cell phones.

It is known that the three dimensional approach does not account for every emotional category, in other words, there are some emotions that cannot be expressed using only the three dimensions. Also, there seems to be some overlap especially in the activation/power domains. However, it seems that it is more feasible to use the three dimensional approach than categorized emotions because there may be a large amount of emotions that are difficult to annotate as categories. Previous work that uses categories use a small set of broad representations of two to five emotions.

2.5 The most important emotional modeling is local, and can be captured by immediate response rules

My last claim is that instead of having to keep track of emotional state throughout the dialog, adequate information can be gathered from looking only at a small window of the dialog. This would allow for simple emotional modeling systems.

Contrary to choosing an optimal dialog act structure, where it is important to know user beliefs and attitudes learned since the start of conversation, emotion modeling requires much less context and can still be useful as seen in [5, 1, 10].

3. A PERSUASIVE DIALOG SYSTEM

In my work in progress, I use the three dimensional approach to detect emotional exchanges present during persuasive dialog (see [1] for more details). A corpus of 10 conversations between a graduate coordinator and students was analyzed. The corpus consisted of a graduate coordinator persuasively informing students about the graduate school option. The corpus was labeled using the three dimensions of activation, valence, and power, each with values from -100 to +100. Correlation coefficients showed some significant relationships between the emotions in coordinator's responses and the emotion in the student's previous utterance. The findings showed that in many cases, if the student sounded positive, the coordinator replied with a positive sounding voice. If the students sounded negative, then the coordinator also sounded negative. The authors also found that there was an inverse relation in the power dimension. If the student sounds dominant, then the coordinator will respond with more submissiveness, but if the student sounded submissive, the coordinator would display more dominance.

Next, more complex immediate response rules were extracted using machine learning methods and slightly better correlation coefficients were achieved. An emotion recognizer was built by extracting acoustic features from the speech signals and using them to predict the labeled emotion levels.

3.1 System realization

Several tools are available to collect speech and recognize both words and emotion in pseudo real-time environments. Here I discuss a dialog system that utilizes a user's voice to assess emotion and spoken words, and responds with persuasive statements. Different from previous work, I plan to automatically choose an appropriate response and render that response with appropriate emotion through prosody.

Figure 1 shows a dataflow diagram for a persuasive dialog system. First the user speaks into a microphone, which is the receptor in, for example, a cellular phone, and the user's words and emotional state are recognized by processing the spoken signal. For recognizing words, pocketsphinx is a viable solution because it is optimized for embedded processors. Emotions are recognized by first computing pitch and energy using the lightweight C program known as dede.

The speech is passed into a dialog manager which will decide on the lexical response. The lexical responses will be retrieved from a database. Several embedded databases packages exist that are fit for mobile devices, e.g. SQLite and VistaDB. This step could also be distributed to a remote server if the database is too large. The emotion is given to the immediate response pattern module, which is a set of switch statements, in order to determine the appropriate emotional color to add to the lexical content based on the user's current emotional state. Both the lexical content and the appropriate emotional coloring will be give to a speech synthesizer capable of producing emotional speech. A possible solution is MaryTTS, which can reside on a remote server and provide synthesized emotional speech as compressed audio.

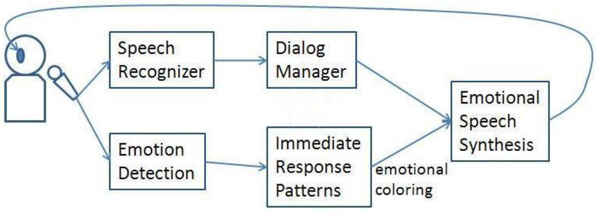

Figure 1: A dataflow diagram for a system that attempts to gain rapport with users

Currently in progress is the evaluation of such a responsive system. Several informal tests have been conducted that suggest that the responsive system is perceived as more adaptive than a baseline system. The persuasive ability of the system will be examined in future work.

4. CONCLUSIONS

I believe that dialog systems should respond to what the user says and also to how the user says it. Emotional state given through prosody is essential to persuasion. Rather than concentrate on methods that use physiological state, face sensors, eye detection, and others, I believe there is valuable data that can be extracted in primary mode of communication in most mobile devices: voice. Mobile devices are at a technological point at which they must utilize emotion as a resource to provide better user experiences.

5. REFERENCES

[1] J. C. Acosta and N. G. Ward. Responding to User Emotional State by Adding Emotional Coloring to Utterances. In *Twelfth International Conference on Spoken Language Processing (in press)*. ISCA, 2009.

[2] P. Andrews, S. Manandhar, and M. De Boni. Argumentative human computer dialogue for automated persuasion. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 138–147, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[3] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber. An emotion-aware voice portal. *Proc. Electronic Speech Signal Processing ESSP*, pages 123–131, 2005.

[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] M. Guerini, O. Stock, and M. Zancanaro. Persuasion models for intelligent interfaces. *Proceedings of the IJCAI Workshop on Computational Models of Natural Argument*, 2003.

[7] I. Mazzotta, F. de Rosis, and V. Carofiglio. Portia: A User-Adapted Persuasion System in the Healthy-Eating Domain. *IEEE Intelligent Systems*, pages 42–51, 2007.

[8] 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.

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

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