A survey of studies of Emotions in dialogue setting

Sigrún Ammendrup

University of Iceland
ammendrup@ru.is

ABSTRACT

When trying to find data for emotional speech there have been different kinds of research methods. Most studies rely on acted emotions, but some studies have taken on different kinds of research methods. In this paper those studies will be investigated as well as studies that have used the method of acted emotions.

1. Introduction

When trying to implement emotions into computers there are many aspects you have to consider. First of all, what kind of a data can you use? Is there a problem getting data for emotion detection and or modeling in dialogue settings? Second problem that arises is what emotions do you want the avatar to express and also how are you going to evaluate how the avatar shows the emotions. It is also optimal that you make research on how the voice depicts certain emotions and how can you do that? Do you research your own face and make the avatar act accordingly or do you make a research to get the widest range of expressions to implement on your avatar. The latter is of course the optimal choice, but then a question arises. How do you make such a research? When trying to find out how people get angry and how they show it in dialogues, you would have to have the people in a lab where there are no distractions so that you can study every movement they make, every eyebrow raise and how their heartbeat rises and what they do with their facial expressions when they are speaking certain sentences or feeling certain
emotions. But how can you get such an accurate measurements as well as get the real emotions that people show when they are angry or happy?

Many studies have been made on emotion detecting in dialogue settings but such researches have almost always been made on people under lab supervisions and the question that then arises is whether or not these emotions are real.

1.2 Emotions in speech
Emotions have a broad influence over behavior. There are some known physical signs where you can see emotions like facial expressions, body language and a certain acoustic features of speech. When it comes to emotions in speech there have been many studies. Some studies suggest that you cannot depict on voice when it comes to emotions since you can have one emotion but several different types of pitch ranges and formants to depict that same emotion.

Some basic emotions that have been studied in speech include:

**Anger**, but anger has been studied well in literature, it has a very broad range in the emotional scale and is therefore too difficult to analyze coherently. Anger can be described as threatening where F0 is tightly controlled and the range is low and near monotone where as when someone is throwing a tantrum has a wide, raised pitch range and therefore totally different from the threatening tone of voice. There for it is impossible to pin down anger as one emotion since it has many sub-types. (Huang, Acero, & Hon, 2001).

**Joy**, when people feel joy their pitch becomes increased as well as their pitch range and the speech rate becomes faster. It has been studied that while smiling the F0 raises and the frequency in the formant also raises. That can be well identified by an untrained listener. (Huang, Acero, & Hon, 2001).

**Sadness**, with most people the pitch becomes normal or lower than usual and that it has a narrower range. The tempo and the rate of the pitch also become slower. Sadness can also be characterized by the way that pronunciation of
words becomes slurred and the rhythm of speech is irregular. (Huang, Acero, & Hon, 2001).

**Fear,** with fear the pitch is characterized as high with a wide range. Rate becomes variable while the pronunciation becomes more precise than usual. The voicing becomes irregular but that can be due to the fact that the respiratory pattern is also disturbed (Huang, Acero, & Hon, 2001).

### 2. Studies on emotions

As previously stated there are problems getting data for emotion detection/modeling in dialogue settings. Anton Batliner has written and talked about this problem in his article *Private emotions versus social interaction: a data-driven approach towards analyzing emotion in speech* (Batliner, 2007). He states that most studies on emotion have dealt with acted emotion, and according to him the only thing you learn by modeling acted emotion is how acted emotion work. To be able to study real emotions you cannot use a model that is based on acted emotions. You will not have learnt anything of any value. Batliner points out that when emotions are being evaluated by humans then it depended on how good an actor the subjects were when it came to evaluating the emotional state being presented. (Chu-Carroll, 2000)

#### 2.1 Computational models of emotion

One computational model of emotion is EMA (Emotion and Adaptation), (Gratch & Marsella, 2004). This application is used so that people can interact with a virtual human using natural language in high stress social setting. High stress social settings can be for instance when you would have to talk to a rape victim and try to get information from them without scaring them or threaten them. When testing this model they wanted to see how people would behave under stressful circumstances and also how their coping skills.

When implementing this research they based their theory on appraisal theory of emotion informed by Smith and Lazarus’ cognitive-motivational-emotive theory which argues that emotion arises from two basic processes: Appraisal and coping (Lazarus & Smith, 1993). Appraisal is the idea that emotions are extracted from our evaluations (appraisals) of events that cause specific
reactions in different people and coping occurs when the person modifies the way they think, for example: employing denial, or distancing oneself from the problem. People may alter the way they think about a problem by altering their goals and values, such as by seeing the humor in a situation.

This research showed a subject a stereotypical episode where the subject’s response was measured several times. Subjects were asked to imagine themselves in an argument, how they would feel and also how they would cope with the situation. Here the study relies on acted and made up circumstances. Their study is still continuing but they show promising results. But since all emotions are acted this study would fall into the category that Batliner points at that you cannot use data that has acted emotions since you will not get accurate results.

2.2 MIMIC: An Adaptive Mixed Initiative Spoken Dialogue System for Information Queries.

The main goal for this system is to enable a spoken dialogue system that simulates human-human dialogue by dynamically adapting dialogue strategies during an interaction based on information that can be detected in the conversation (Chu-Carroll, 2000). Since this system is a telephone-based dialogue system it only has to recognize if the human is getting annoyed or if the system is not finding the right answers to the human’s question. The main components for this system are: Telephony server, speech recognizer, NLP/Dialogue component and Text-to-speech engine. The system was evaluated in a controlled environment where eight users interacted with the system and the user satisfaction was assessed by asking the subjects to fill out a questionnaire after interacting with it.

The system was trained on a corpus that had transcribed dialogues where each turn was annotated with task/dialogue initiative holders and the cues of the observer.

This system has a lot of possibilities since it can be built upon to have more annotations and therefore make it more reliable.
2.3 An Acoustic Study of Emotions Expressed in Speech.
This research was aimed at finding different acoustic properties of speech while a person was feeling four different emotions. These emotions were sadness, anger, happiness and neutral. The aim was to find out how speech is modulated while the speaker’s emotion changed from neutral to a certain emotional state. They measured acoustic parameters related to speech prosody, vowel articulation and spectral energy distribution. This entire research is based on an actress, acting these emotions and all the conclusions from this study come from studying the changing of acted emotions (Yildirim, 2004).

2.3 Annotating Emotion in Dialogue
This research was built up on seeing how dialogue is affected by participant’s emotion and how that could be implemented in Human-Machine dialogue. They wanted to see if it annotating emotion in dialogue that has previously been annotated in other ways would reveal useful correlations that have not been seen before. They used a corpus which is built up on 37 dialogues between nurses and patients. These dialogues are built up on real conversations and therefore contain real emotions. This is the perfect corpus if you want to study realistic dialogue (Craggs & Wood, 2004).

2.4 Emotion Estimation Algorithm Based on Interpersonal Emotion
To be able to detect emotions in dialogue you have to have emotionally tagged corpus. The research Emotion Estimation Algorithm Based on Interpersonal Emotion did that. They proposed a method of emotion estimation by putting different kinds of weight on words depending on the emotion that they are showing, and also having interpersonal relationship expected from sentence. They programmed the system to find out whether or not the words describing emotions were belonging to the person speaking them or whether or not they were used to describe the feelings of someone else (Matsumoto, Ren, Kuroiwa, & Tsuchiya, 2007).

To be able to have a corpus to practice the system on, they built an originally created Emotion Word Dictionary. It is built up on content words taken from
sentences that are expressing emotions. The words have different kinds of weight depending on the emotion that they are describing.

1190 sentences were tagged with 9 emotions and then they had a smaller corpus of 90 sentences tagged with 9 emotions that was not used to train the system. Both corpuses were tagged by one person, and after the system had tagged unseen sentences the results was compared to the manually tagged sentences (Matsumoto, Ren, Kuroiwa, & Tsuchiya, 2007)

This research was made for the Japanese language. For Scandinavian languages this could proof to be troublesome because of word ambiguities. In Icelandic for example the word for something beautiful could be the same as when you are describing something very ugly.

2.5 Design of a Hungarian Emotional Database for Speech Analysis and Synthesis
A database is being built in Hungary for emotional speech database. In the article Design of a Hungarian Emotional Database for Speech Analysis and Synthesis the work is being described. They are focused on a core set of 12 emotions as well as collecting and annotating children tales. For the collection of the database they would get 4-8 actors to read sentences which were judged as neutral to see the difference between them and sentences that were judged as being biased, or emotionally charged (Fék, Németh, Olaszy, & Gordos, 2004).

They would then get people to judge what emotion is being portrait in different sentences to get accurate measurements of how emotions are being displayed in the sentences (Fék, Németh, Olaszy, & Gordos, 2004).

The research has not been completed so what is being described in the article is how they would like to implement the research.

2.6 The Vera Am Mittag German Audio-Visual Emotional Speech Database
This corpus is based upon 12 hours of recording from a German TV talk-show which is called „Vera am Mittag‘. The recordings are based on broadcasts, dialogues and utterances. Everything has been imported into the corpus. What
makes this very special is the fact that the speech that has been gathered is both spontaneous and very emotional as well as being unscripted since these are authentic discussions between guests of the show and the talk-show host.

What makes this corpus unique is the fact that you have speech, emotions and visual clues as to what is going on all at the same time (Grimm, Kroschel, & Narayanan, 2008). The only problem with this corpus is the fact that you cannot control the situation around which emotion is being displayed at a certain time.

From this information they made the VAM corpus which is split into three different modules, the video, audio and facial corpus. This corpus is available for researchers that would like to annotate or use the corpus to investigate emotions in dialogues.

### 3. Problems with speech corpora

Within this field the biggest problem is not what to study or how to study emotions in speech but how to get the appropriate material to use in your research. When analyzing emotions in dialogues you want to have materials that show real emotions in the speech and in the dialogue that you are studying.

When it comes to researching emotions there have been three main research methods:

- The first one, which is the most common one, is when actors deliver lines and act emotions and express different emotions. Then the accuracy of the data is reliant to which of the emotion is being acted. Since the emotions are acted it is therefore likely that pronunciation and emotions are made to be much more exaggerated then they would be in a normal setting.

- The second research method is when participants are provoked into a certain emotional state and then their speech is recorded. Then the emotion is induced. It would be possible to induce dialogue while the
person is experiencing a certain emotion, but that would be very inaccurate because it would be unlike real conversations.

- The last method would be to just record people while they are unaware of it. Then you would get genuine speech with real emotions which would represent human behavior. This method has been used in a research before in 2000 which was called Criteria for emotion recognition from verbal and nonverbal expression: studying baggage loss in the airport (Scherer & Ceschi, 2000). The problem with this method is the fact that it is illegal to record people speaking to someone else without their knowledge. There are many ethical questions and concerns.

4. **Realistic solution to finding data for emotional speech corpora**

The most desirable corpus to research emotions in dialogs would be a corpus that had genuine conversations where people would express emotions, or have some emotions about the topic that they are talking about. Of course people would have to be oblivious to the fact that they were being listened to or let alone recorded.

When trying to induce emotions in people there is likelihood that the dialogue between people would become unnatural or strange and therefore not usable for research. The way that you could achieve the goal of having natural dialogue where you would have emotions would be where people were put in unusually emotional circumstances, but without them knowing that they were being watched.

The solution to this problem is not simple. The research must always follow certain rules and be ethical. My proposal would be to get a group of people to solve certain problems, where they would sign a paper that everything that they do or say can be used for research purposes, but tell them that the research is about how well people work together under stressful circumstances. Not to tell them that the main focus would be on their dialogue. Then you would have to have them in the experiment for a long time to get them to know each other and to start talking freely about, both the assignments as well as daily life. The optimal conclusion would be if they would
start to talk and behave as they would normally do, that is showing their true self. Every part of the lab would have recorders so that when people would go get some coffee or relax between assignments you would record their dialogue.

Even though the corpus VAM (Vera Am Mittag) is not recorded under supervised lab conditions the fact remains that there you have corpus of spoken dialogues with emotions and facial expressions. The only thing that could corrupt this data is the fact that people are very well aware that they are being recorded and that they will appear in television so they might not behave exactly the way that they would usually do. But in my opinion that is more realistic then getting actors to quote some lines with emotions that they believe are representing the feeling that they should be showing.
Bibliography


