Affective Speech for Social Communication: Implementation Challenges in Text-to-Speech for Short Messages

Alia Amin [1]
User System Interaction
Technische Universiteit Eindhoven
Den Dolech 2,
5612 EZ, Eindhoven
The Netherlands
alia.amin@cwi.nl

Jimmy Wang
Industrial Design Dept
Technische Universiteit Eindhoven
Den Dolech 2
5612 EZ, Eindhoven
The Netherlands
c.wang@tue.nl

1. Author's current affiliation:
Semantic Media Interfaces, CWI,
Amsterdam.

CHI 2006, April 22–27, 2006, Montreal, Canada.
ACM 1-xxxxxxxxxxxxxxxxxx.

Abstract
The flexibility to choose from different modal content presentation will be an important feature in future ubiquitous application. Currently, short messages (e.g. SMS/MMS) are only available in visual form. However, in certain situations, users may like to have these messages presented in audio form. We explored the alternative of presenting short messages in affective synthetic text-to-speech form special for social communications between teens. Evaluation of this alternative presentation reveals that, for emotion recognition, it was easier to interpret emotion messages generated from affective synthetic speech. Although there is no actual difference in the way people think they were able to derive emotions from both types of messages. For teens, affective synthetic speech is sometimes fun to use.

Introduction
The flexibility to choose the modality in which communication messages is presented, i.e. visual or audio, is an important feature in truly mobile and ubiquitous applications. Short messages are usually delivered in text form. However, there are situations where information is best presented in audio form. People who are busy with an activity, e.g. driving a car or riding a bike, would prefer hands-free and eyes-free

interaction. Another example is people with visual or motor disabilities. These people need to have text information be delivered in an audio form rather then visual form. Furthermore, speech is the most natural means of communication which has existed as long as human interaction.

Applications using synthetic speech are still perceived as "cold" and "flat" [2]. Even though flat synthetic speech is relatively common in an interaction with a machine, it is not desirable to have "flat" synthetic speech for all type of applications e.g. story telling and human-human communication via technology (email, SMS, chat). One way of making synthetic speech less "flat" is by adding emotion to the speech. In our project, we want to find out if users can derive the emotion from a short message through affective synthetic speech.

Deriving emotions from a text message

To convey affective messages, the mobile phone should be able to recognize emotions. This is done by getting the emotions from the emoticons in the message. This solution was inspired by the fact that in SMS, people have solved the problem of lacking emotions by using emoticons in addition to the text. In short messages applications, conveying emotions to text is possible with the help of emoticons. For example, the message "It is my birthday today :-)" infers that the sender is happy about the situation, but the message "It is my birthday today :-(" infers that the sender is unhappy about the situation. By deriving emotions from the emoticons in the messages, it is hoped that affective synthetic speech can enhance the understanding of the psychological (sender's emotional state) of the messages for at least six basic emotions purposed by Ekman and Friesen (1976): happy, sad, fear, angry, disgust and surprise. There are difficulties in embedding these emotions in a synthetic speech because there are many variables that play a role in forming a speech waveform. These difficulties are discussed in the following section.

Related Work in Affective Speech

Mimicking natural spoken language in synthetic speech is not a trivial task. Synthetic speech tends to be perceived as unnatural, in other words; it lacks emotions [2]. In normal speech people get cues from the voice of the speaker, the so-called paralanguage (e.g. intonation, melody, pitch and pace). From these cues people infer the emotional state of the speaker and the intention and meaning behind the message. Research on expressive synthesized speech has been done by Cahn (1989) for the English language[2]. Through a real time manipulation of the affect parameters by a certain (emotion) linguistic rule, an expressive synthesized speech can be produced. In natural speech, people convey emotions by means of prosody. Such prosodic features are pitch, speech rate, rhythm and loudness [1,6,7].

Research Method

Generating Affective Speech

Our participants are native Dutch speaker, thus, we generated synthetic speeches using the Dutch diphone synthesizer called Spengi developed by IPO[1] and create an affective speech effect with a wave editor called GYPOS. We decided to recreate the synthetic speech messages by mimicking three most important variables: the pitch, speech rate and loudness of a real voice [1,2]. This can only be done non real-time with

[1] http://www.let.uu.nl/~audiufon/data/difoon.html

1. "Hey ik ga vanmiddag nog ff tilburg in met wat vrienden dus ben wel rond het avond eten thuis mzzl" (translation: I am going to Tilburg with friends and will be home in the evening for dinner)

2. "gvd waar, waar de gister ik stond daar gewoon alleen in v'waard echt een naaistreek" (translation: Goddammit, where were you yesterday, I was alone in v'waard, it is really terrible)

3. "ey aikol je bent toch jarig 2day?! Nja anyway happy bday ook al zit je leve tege x djessie". (translation: Hey Aikol, its your birthday 2day right?! Well, anyway happy bday, although your life sucks, kiss djessie)

figure 1: Samples of short messages given by the teens

GYPOS. Afterwards, we conducted two user studies: informal interviews with teens and questionnaires with young adults to measure whether emotions can be interpreted correctly through affective synthetic speech.

The Interview

In total 8 teenagers, five girls and three boys, with ages from 16 to 18 were interviewed. Prior to the interview, the participants were asked to give samples of SMSs which they had made themselves. We converted these text messages to affective synthetic speech messages. For every SMS, there were several samples presenting all basic emotions (e.g. sad, happy, angry, etc.) and one "flat" (no emotion). In the interview, the teens were asked to listen to the samples of other people's messages and interpret the emotions. We also showed them an SMS on a mobile phone and asked them to say which ones they preferred the most and why. The results of the interviews reveal that it is difficult to identify emotions such as disgust, fear and surprise in the synthetic speech. Participants also said that the female synthetic voice is not suitable for the message from a male. During the end discussion the teens said they found the affective synthetic speech fun to use for certain messages.

In the second user study, six people participated. Three males and three females with ages from 24 to 30. The purpose of the experiment was to evaluate the emotions behind the messages. Each participant receives three affective messages (figure 1) and had to guess which emotions they entail. The result was in comparison on how well subjects were able to infer the emotion from the plain text versus the affective synthetic version. The questionnaire's average scores of how well people thought they were able to derive the emotion from the two messages were calculated (1–very bad, 5–very good). The results were that people gave an average of 3.77 for the text messages and an average of 3.61 for the affective speech messages. Thus, no actual difference was measured. However, when comparing the actual emotion that was intended for the messages, it showed 50% correct interpretations for text the text messages and 67% correct interpretations for affective speech messages. This means, although people say it is easier to derive emotions from text messages, they scored better with affective speech messages. In general, people found it easier to infer emotions from the affective speech messages when the words used was quite neutral. However, if the content contained obvious words, like swearing, it was easier to infer the emotion from the text message. An interesting finding is that the pitch of the female synthetic speech used in the messages was quite high and this confused people in thinking that the voice message was angry.

Discussion

Affective speech is supposedly intuitive and simple for the sender because it does not require him to learn new syntax. However, based on our research, from the receiver's perspective, affective speech can add another dimension but **not** replace the current text message for several reasons:

1. Emoticon lexicon is richer than expressive speech. During the past years of the Internet boom, emoticons have developed into rich variations. There are many well known emoticons today and the list is still growing [3]. These rich expressions can not be expressed with natural speech, e.g. :D (big smile) and B) (I am with glasses smiling). On the other hand, the basic emotion of fear is not represented in emoticon language. The intention of fear is usually expressed in the words itself

or might be inappropriately represented in another form of smiley (e.g. I am afraid my boss will fire me :( ) non the less, the intention is usually perceived correctly;
2. Abbreviation and language style: Along with emoticons, abbreviations are also dynamically changing. Examples in Dutch language are "ff" and "mzzl" (see figure 1). We found there are abbreviations that are specifically used by certain age groups; e.g "xje" for kusje (kiss) and "kben" for ik ben (i am). Teenagers are very creative in shortening their text messages and often have their own style in text which makes it difficult to translate to synthetic speech e.g. "Hoiiiii Aikol" (redundant letters) or "2day" (using numbers instead of letters).
3. Generated speech should correspond to the gender. Otherwise it would be perceived as strange or even inappropriate.
4. A neutral plain synthetic voice is often interpreted as being angry. This could be explained as the behavior of a person talking plain and short when he/she is angry. This is also the case for high pitch voice.
To properly implement this affective speech, a solid linguistic knowledge is required. The database of abbreviations, emoticons and symbols used in text messaging and the algorithm for different emotions in synthesized speech have to be developed to be able to properly translate these emoticons into affective speech. Part of the research on this topic is carried out by some researchers [2][6]. Finally, from the evaluation we know that emoticon lexicon is very rich. There might be some emotions or expressions which can only be expressed by emoticons and vice versa there can be some emotions which are better expressed through speech. In the end, there should be flexibility to choose from the possible modality to interact with. If all barriers mentioned above can be solved, using affective speech in mobile devices may just be something that is just around the corner.

Acknowledgements

Thank you to Jacques Terken for his guidance in this research and to other project members: Bram Kersten, Elly Pelgrim, Olga Kulyk.

References

[1] Cahn, J.E., Emotional & Expressive Synthesized Speech, (1989), retrieved at May 2005, from http://alumni.media.mit.edu/~cahn/emot-speech.html.
[2] Cahn, J. E., Generation of Affect in Synthesized Speech. Proceedings of the 1989 Conference of the American Voice I/O Society. Newport Beach, California. September, 1989, 251-256.
[3] Emoticons & Smilies Page, retrieved at May 2005, from http://www.muller-godschalk.com/emoticon.html
[4] Examples of Synthesized Speech, retrieved at May 2005, from http://www.ims.unistuttgart.de/~moehler/synthspeech/examples.html
[5] Lee, C.M. & Narayanan, S., Towards detecting emotions in spoken dialogs., (2004), retreived May 2005, from http://sail.usc.edu/~cml/final.pdf.
[6] Mozziconacci, S.J.L., Speech variability and emotion: production and perception, Technische Universiteit Eindhoven, Eindhoven, 1998.
[7] Walker, M.A., Cahn, J., and Whittaker, S.J., The Improvisation of Linguistic Style: Social and Affective Bases for Agent Personality. Proceedings of the ACM Agents '97 Conference. February, 1997. 10-17, retrieved at May 2005, from http://alumni.media.mit.edu/~cahn/autonomousagents.html