An Integrated Approach to Emotional Speech and Gesture Synthesis in Humanoid Robots

Philipp Robbel
Personal Robots Group
MIT Media Laboratory
77 Massachusetts Avenue
Cambridge, MA, USA
robbel@mit.edu

Mohammed E. Hoque
Affective Computing Group
MIT Media Laboratory
77 Massachusetts Avenue
Cambridge, MA, USA
mehoque@mit.edu

Cynthia Breazeal
Personal Robots Group
MIT Media Laboratory
77 Massachusetts Avenue
Cambridge, MA, USA
cynthiab@media.mit.edu

ABSTRACT

This paper describes an integrated approach to recognizing and generating affect on a humanoid robot as it interacts with a human user. We describe a method for detecting basic affect signals in the user's speech input and generate appropriately chosen responses on our robot platform. Responses are selected both in terms of content and emotional quality of the voice. Additionally, we synthesize gestures and facial expressions on the robot that magnify the effect of the conveyed emotional state of the robot. The guiding principle of our work is that adding the ability to detect and display emotion to physical agents allows their effective use in novel application areas such as child and elderly care, healthcare, education, and beyond.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics—*speech and gesture generation*; H.1.2 [**Models and Principles**]: User/ Machine Systems—*affect recognition*

General Terms

Human factors, Algorithms

Keywords

Human-robot interaction, speech and gesture generation, affect recognition

1. INTRODUCTION

The research described in this paper aims at creating a physical agent that is capable of enganging in convincing interactions with a non-expert human user. Although significant advances have been made in machine learning theories and techniques, existing frameworks do not adequately consider the human factors involved in developing robots that learn from people who lack particular technical expertise.

To make progress, we need to better understand the human factors associated with teaching robots. We believe that in order to enable applications in healthcare, elderly care, and education (among others), an effective agent should be able to detect affect signals in the user's speech and generate responses appropriate to the emotional state of the user. This paper summarizes our initial efforts in this area. We contribute an integrated mechanism for generating gestures and emotional speech on a humanoid robot platform in a dialog setting during which we keep track of the emotional state of the user. Our platform is the MDS robot as well as a virtual model thereof, as presented briefly in the following section.

1.1 The *MDS* Robot Platform

The mobile, dexerous and social (*MDS*) robot is the latest humanoid robotic platform at the MIT Media Laboratory's *Personal Robots Group* and has been in use since 2007. The purpose of this platform is to support research and education goals in human-robot interaction, teaming, and social learning. The robot comes on a mobile, self-balancing base and is equipped with two arms, a flexible torso and a high degree-of-freedom face supporting the display of a range of facial expressions. With respect to size, the robot roughly matches that of a 3-year-old child. Sensors on the robot include a color CCD camera in each eye, an indoor active 3D IR camera in the head, four microphones to support sound localization, and a wearable microphone for speech capture. A picture of the MDS sociable robot is shown in Figure 1.

The MDS robot allows for expression of a great range of affect through gestures and facial animation. An early video demonstrating the expressiveness of our robot has been available on YouTube since April 2008 and has been viewed more than 420,000 times[1] since then.

For purposes of this paper, we assume a simple dialog scenario where a short speech segment of the user is followed by a reaction of the robot that should be convincing both in terms of the voice quality and the animation of the robot. In order to implement this system, we address three main areas that are presented in more detail in the sections below. First, we have to implement a reliable way to detect *the user's emotional state*. We restrict this part to the interpretation of speech features that we extract directly from the input into the microphone. Other ways to interpret the affective

AFFINE '09, November 6, 2009 Boston, MA
Copyright 2009 ACM 978-1-60558-692-2-1/09/11 ...$10.00.

[1] See `http://www.youtube.com/watch?v=aQS2zxmrrrA` for the official MDS robot video on YouTube.

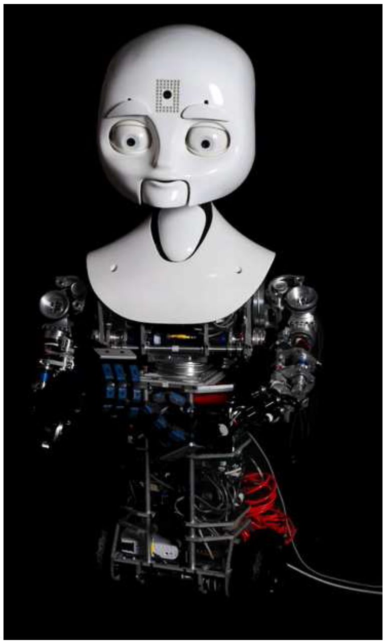

Figure 1: The MDS sociable robot.

state, such as facial expression of the user, are presently not considered but could be added fairly easily to our overall architecture. Second, we have to *synthesize a response to the user in a timely fashion.* This response must appear to be correct both in terms of speech content and in terms of the affective quality of the voice. A rule-based system determines which emotional coloring to choose. Lastly, the speech has to be played back on the robot in a convincing manner. This involves choosing an appropriate robot animation and to synchronize the generated speech with *gestures and mouth motion of the robot* (lip-syncing).

2. EMOTIONAL FEATURE EXTRACTION AND SELECTION

This section provides a brief overview of the data collection methods, feature extraction and selection process. The feature selection algorithms yield a set of prosodic features that are believed to be the correlates of two particular mental states that we are interested in and reply to: *excitement* and *boredom.* For purposes of this paper, we collected a small database of short and medium utterances produced by a set of actors with feigned emotional patterns. There were 10 subjects in this study, of which 50% were male and the remaining 50% were female. In total, 80 utterances were produced by the actors which were split evenly between *excited* and *bored* speech.

2.1 Speech Feature Extraction

To compute the prosodic information from spoken utterances, features related to segmental and suprasegmental information, believed to be correlates of affect, were calculated. Computed features were utterance level statistics related to pitch [10, 2], energy, pause patterns, and speaking rate.

The moving patterns of pitch in speech often provide useful information not available through any other communication channels. For example, abrupt changes of pitch patterns are correlated with frustration or anger. Low average pitch value and slightly narrower pitch range of an utterance indicate disappointment [7]. Rising patterns of pitch tend to correlate with questions or curiosity, whereas falling patterns of pitch indicate a belief or statement. Therefore, high occurrences of falling edges in an utterance are very likely to fall into the category of an instruction or explanation.

Energy flow (correlated with the sensation of loudness) and pause patterns are also potential candidates for modeling affective states. For example, there are more possibilities for long pauses in an utterance when one is explaining or instructing rather than providing a firm reply or acknowledgement.

In total, we compute a list of 36 features for each segment of speech collected in the training database. Among these are minimum, maximum, mean, mode and standard deviation of pitch and intensity as well as quantifications of the edge contours (magnitude of highest rising and falling edges and their numbers). Statistics over formants (corresponding to the resonance frequencies of the vocal cords) and the duration of the speech acts are also taken into account. Throughout this work, we used Praat[2], an open source software for acoustic analysis, to generate the initial set of features.

2.2 Feature Selection

In order to find out which feature sets carry the most information when classifying between bored and excited input speech, we investigated different feature selection and evaluation techniques with the Weka data mining package[3]. Candidate feature subsets were generated through greedy stepwise search and evaluated with the CFS subset evaluator [5], consistency subset evaluator [6], and chi-squared attribute evaluator. For our purposes we found that the *mode of the pitch*, *mode of the intensity*, and the *difference between maximum and minimum intensity* contributed to most of the variability that was needed to differentiate between the two emotional states. The final classifier employed in this work was a decision tree with learned thresholds for each of these variables. In our experiments with a microphone at a fixed distance from the user we achieve roughly 80% prediction accuracy.

3. EMOTIONAL SPEECH AND GESTURE SYNTHESIS

The second component of our system generates emotional speech on the MDS robot. Our goal is to generate appropriate responses to the user, both in terms of content as well as pitch and emphasis. For that, we employ a text-to-speech (TTS) system to synthesize a number of sentences with different emotional styles. In the current system, we employ a state machine-based dialog system to select sentences that drive the dialog. Sentences are selected based on the emotional state of the user as detected from the user's latest utterance. At present, our robot inquires about the day of the user and either selects emphatic responses if the user appears bored or mirrors the user's emotion if it is classified

[2] Available on http://www.fon.hum.uva.nl/praat

[3] See http://www.cs.waikato.ac.nz/ml/weka/

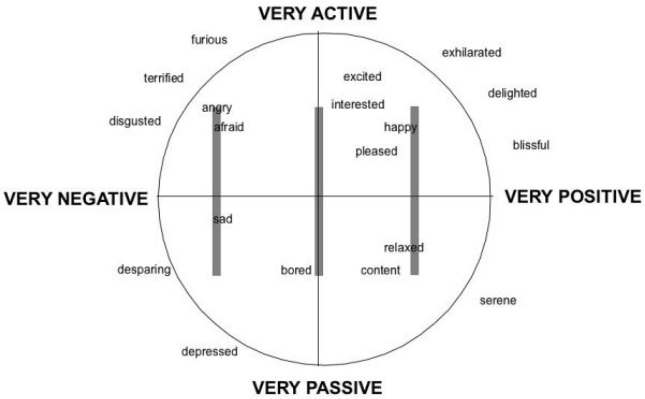

Figure 2: The emotional spectrum of Plutchik [8] and the areas that CereVoice attempts to simulate as gray bars. Retrieved from [4]. See [9] for a similar circular model of affect.

as excited. The goal of this robotic agent is to drive the user away from a bored to an excited emotional state.

Most commercial systems come with neutral voices that are not specific to any domain or emotion and not directly usable for the expressive synthesis we would like to achieve. Recent research has made some advances in this direction by allowing the user to select among different pre-recorded voice databases (e.g. for “happy” or “calm” speech) and to apply digital signal processing (DSP) to further refine rate or amplitude of the output signal [4]. The current state-of-the-art emotional speech synthesis is semi-supervised through explicit annotations to the textual input before synthesis. Our project attempts to *infer these annotations* automatically in a dialog situation with the user. Furthermore, we not only generate emotional speech but also appropriate robot gestures and facial expressions that emphasize the particular emotion. For the TTS part of this work, we settled on the commercially available CereProc speech synthesis SDK [1].

CereProc supports emotional annotations through standard XML tags in the input text; competitive products offer similar annotation features. Examples for those are database selection tags (for “stressed”, “calm” or “joking” genres) and DSP annotations to modify amplitude and speech rate. In Figure 2, a general idea of the emotions that the synthesizer tries to cover are provided. The distinction across the negative-positive continuum is roughly covered by the database selection whereas the transition between active and passive is achieved through DSP modifiers. In practice, the effects of synthesized emotional speech can be underwhelming due to the lack of refined speech databases and because too aggressive DSP introduces artifacts which can render the output to sound unnatural.

We believe that adding further cues, such as robotic facial expressions and gestures can remove some of these problems associated with state-of-the-art emotional speech synthesis.

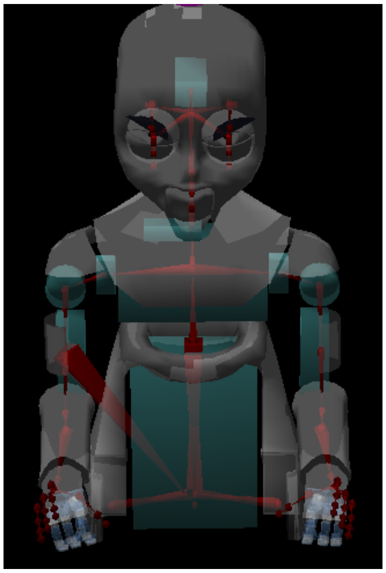

Figure 3: A single frame of the “angry” animation blended with the generated speech and visualized in our robot simulator. Clearly visible is the head posture and the characteristic eyebrow location in the robot skeleton (shown in red).

3.1 Real-time Gesture Synthesis

Our robot architecture, based on [3], allows a humanoid robot to play back and adjust the timing of custom animations involving all joints and to blend seemlessly between them. For our demonstration, we recorded “happy”, “angry” and “neutral” animations of the robot off-line, involving most joints of the robot. These animations have a large effect on the robot posture, its facial features (such as eyebrows) and its arm motion. All animations were created by a professional animator by hand and exported from Autodesk Maya into our robot animation toolchain that maps the joints of the virtual robot model to the robot motors. We generate accurate lip-syncing by creating mouth animations on the fly, based directly on the phonemes extracted by our speech synthesis system. This is achieved by blending the MDS mouth animations for the different basic sound profiles in short succession. The convincing results can be played along with the audio in simulation and on the robot in real-time. Figure 3 shows a single frame of the “angry” animation in our simulator.

The final system is able to convincingly blend robotic gestures with spoken text—emotional quality of the voice is chosen automatically and augmented by the appropriately selected gestures and facial features.

4. CONCLUSIONS

In this paper we presented our system for generating affective speech on the MDS robot platform. We demonstrated that for two affective states of the user (*excited* and *bored*), we can generate meaningful responses with high accuracy. For now, emotional responses may well have been pre-recorded but our system allows for generating custom messages on the fly. In the future, the dialog may contain the user name, for example, or depend on other external information (such as the weather or the stock market) that we could easily integrate into our text-to-speech architecture.

Future extensions are therefore mainly concerned with creating a more sophisticated dialog experience with the user. Other extensions to the described architecture include refinements to our emotion classification method to make it more robust to changing microphone distances. We would also like to include other means to detect the emotional state, e.g. based on visual feedback on the user's facial expression. Future work will also include a user study comparing the effects of emotional speech synthesis with and without the inclusion of robotic facial expressions and gestures. This study will not only look at how good the classification of the human emotional state is but also quantify the degree to which human users can infer the intent of the robot during the dialog.

5. REFERENCES

[1] M. P. Aylett and C. J. Pidcock. The CereVoice characterful speech synthesiser SDK. In *IVA*, pages 413–414, Berlin, Heidelberg, 2007. Springer-Verlag.

[2] R. Banse and K. Scherer. Acoustic profiles in vocal emotion expression. *J. Personality Social Psychology*, 70(3):614–636, 1996.

[3] B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. P. Johnson, and B. Tomlinson. Integrated learning for interactive synthetic characters. In *SIGGRAPH*, pages 417–426. ACM, 2002.

[4] Cereproc Ltd. Cereproc research and development. `http://www.cereproc.com/about/randd`, August 2009.

[5] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, 1999.

[6] H. Liu and R. Setiono. A probabilistic approach to feature selection - a filter solution. In *ICML*, pages 319–327. Morgan Kaufmann, 1996.

[7] R. W. Picard. *Affective Computing*. The MIT Press, Cambridge, Massachusetts, 1997.

[8] R. Plutchik. *The psychology and biology of emotion*. Harper Collins, New York, 1994.

[9] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[10] C. Williams and K. Stevens. Emotions and speech: Some acoustical correlates. *J. Acoustic Soc. of Am.*, 52(4):1238–1250, 1972.