5ZA-4

The DESIRE Model: Cross-modal emotion analysis and expression for robots

Angelica Lim **Tetsuya Ogata** **Hiroshi G. Okuno**
Graduate School of Informatics, Kyoto University

1 Introduction

In Japan's aging society and abroad, interviews reveal that the elderly consider communication and social contact as important as their health [1]. Conveying emotion clearly is an important part of social communication [2]. Since care robots for the elderly are becoming more popular, our work focuses on making robots understand and express emotion for better quality of life.

In this study, we develop and test an emotion framework that encompasses both voice and gesture for humanoid robots. Until now, robot emotion studies have been limited to one modality at a time. For example, Laban Movement analysis [3] was used for robot movement, and DECTalk was used to add affect to Kismet's voice [4]. We attempt to find a general approach or "underlying code" for emotion focusing on *how* one moves, or *how* one speaks. The eventual goal is to analyze and convey emotions in a rich, multi-modal manner using one unified emotional framework. This framework would be useful not only for multiple modalities, but arbitrary robot forms, such as robot vacuum cleaners [5]. In this paper, we propose such an emotion framework, and test it by implementing an emotion transfer system, which converts emotional voice to robot gesture.

2 An Emotion Transfer Framework

We propose a framework (Fig. 1) that models emotion through dynamic parameters of speed, intensity, regularity and extent. For short, we call this parameter set **DESIRE: Description of Emotion through Speed, Intensity, Regularity and Extent**, or simply **SIRE**. Speed and extent have been widely accepted in the Human-Robot Interaction (HRI) community to convey some aspects of emotion [3] [5], and here we study two other parameters called regularity and intensity. Our hypothesis is that certain values of SIRE underlie the same emotions in voice and gesture. An extension to music and a full description of the approach is discussed in [6]. In short, the DESIRE framework consists of:

1. *Dynamic parameters*, representing universally accepted perceptual features relevant to emotion (SIRE). We define them as a 4-tuple of numbers $S,I,R,E \in [0,1]$.
2. *Parameter mappings*, between the dynamic parameters and robot-specific implementation.

The parameter mappings can be divided into two layers as shown in Figure 1: (1) a *hardware-independent layer* and (2) a *hardware-specific layer*.

2.1 Hardware-independent layer

The DESIRE framework was inspired by commonalities found between emotion in movement, voice and music [7] [8]. For example, speed is called *rate* in speech literature [9] or *velocity* in gesture [12]. We have summarized our review in Table 1.

2.2 Hardware-specific implementation

We provide here the mappings shown in Fig. 1 for 1) extracting SIRE from emotional speech audio samples, and 2) generating motions from SIRE on the NAO Humanoid robot.

2.2.1 Extracting SIRE from Voice

The studies in Table 1 provide a good theoretical basis for how to map voice to SIRE parameters. In this section, we assume an input speech sample $x(t)$ with sample rate f_s and length N. In our experiments, these result from audio files recorded at 16kHz.

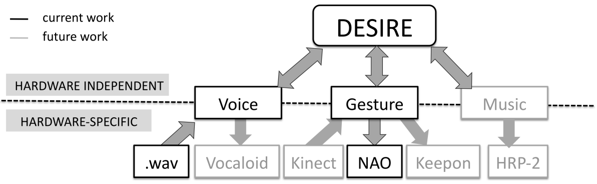

Figure 1: Overview of DESIRE cross-modal emotion transfer framework.

Speed is mapped here to speech rate, or more specifically, syllables per second. For the purposes of this study, we manually provide the number of syllables b. We assume that the sentence sample is clipped at the beginning and end of the utterance, giving us $b * f_s/N$ syllables per second.

Intensity is implemented here as voice onset rapidity. More specifically, we find the power trajectory $p(k)$ of $x(t)$ and calculate its maximum rate of change. The power is given for every frame of size n (in our experiments, $n = 1024$) by $p(k) = \sum_{i=0}^{n-1} x(k \cdot n + i)^2$, and onset rapidity is $\max_{k=1,\ldots,N/n}(p(k) - p(k-1))$.

Regularity is mapped here to the inverse of jitter in the voice sample, as jitter has been related to vocal "roughness" in [10]. Jitter is defined for each utterance as $1/(N-1) \sum_{t=1}^{N} |x(t) - x(t-1)|$.

Extent is defined as the range of pitch in the speaker's voice. We used the Snack sound toolkit[3] implementation of the average magnitude difference function (AMDF) [11] to extract the utterance's F0 trajectory, taking extent as the difference between the lowest and the highest F0's.

Scaling was performed in a similar fashion for all of SIRE. Given the minimum and maximum values for each parameter (experimentally chosen), we linearly scale to achieve a parameter between 0 and 1. For instance, pitch range was linearly scaled between a minimum F0 of 40 Hz and a maximum F0 of 255 Hz. As for speed, we used a minimum speech rate of 2 syllables per second and a maximum speech rate of 7 syllables per second. In future work, we should study how this could be adapted to the speaker, for example by defining extent as the user's deviation from their pitch average.

2.2.2 Gestural mappings for NAO Humanoid

In this section we briefly describe how we implement the perception of speed, intensity, regularity and extent on Aldebaran Robotics' humanoid robot NAO[2]. A gesture is considered here as a simple motion from a "base posture" to an "extended posture" and back to the "base posture", each reached at target times t_0, t_1 and t_2, respectively. Figure 3 shows example postures for arms; we define head gestures similarly.

Speed is mapped by performing a simple linear down-scaling of all target times for higher speeds. *Intensity* is increased by

[3] www.speech.kth.se/snack/ [2] www.aldebaran-robotics.com

Table 1: DESIRE parameters and associated emotional features for modalities of voice, gesture. Features in *italics* were used in our study.

Parameter	Description	Voice	Gesture
Speed	slow vs. fast	*speech rate* [9], pauses [7]	*velocity* [12]
Intensity	gradual vs. abrupt	*voice onset rapidity* [7]	*acceleration* [12], power [13]
Regularity	smooth vs. rough	*jitter* [7], voice quality [9] [7]	*directness* [12], *phase shift* [14] [15]
Extent	small vs. large	*pitch range* [9], loudness [7]	*spatial expansiveness* [13], contraction index [12]

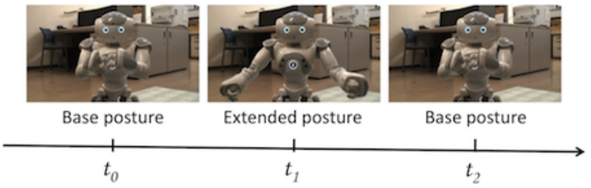

Figure 2: Timeline of an arm gesture.

Table 2: Sequences with best agreement between evaluators and their corresponding SIRE values.

Emotion	Agreement (%)	S	I	R	E
Happiness	60	0.72	0.20	0.22	0.74
Sadness	75	0.12	0.44	0.71	0.42
Anger	60	0.58	0.92	0.24	0.9
Fear	65	0.93	0.72	0.34	0.47

bringing t_0 and t_1 temporally closer together, effectively increasing the relative acceleration to reach the extended posture. *Regularity* is implemented either as joint phase shift and directness, which can be thought of as temporal and spatial regularity, respectively; for arms, a more irregular movement is created by temporally "shifting" one of the arm movements, and for the head, an irregular movement is created by adding side-to-side movements. The amount of side-to-side movement is determined by a random variable taken from a normal distribution with variance inversely proportional to R. Finally, *extent* is calculated by updating the effector's extended position, scaling it linearly between the base and extended positions depending on the value of E.

3 Evaluation

We recruited 20 evaluators from Kyoto University Graduate School of Informatics. As input, we used 16 audio samples taken from the Berlin Database of Emotional Speech[4], which is a database of emotional speech recorded by professional German actors. Each sample was a normalized wave file at 16 kHz, 1.5 to 3.9 s long, all of the same sentence. Four samples each of happiness, sadness, fear, and anger were used, all with recognition rates of 80% or higher by German evaluators.

Given SIRE values extracted from these audio samples, we generated 16 movement sequences using a simulated NAO shown on a projected screen. Only one type of gesture was shown (an extension of both arms in front of the robot), repeated four times in series for each sequence. After each sequence, the participants chose one of happiness, sadness, anger, or fear in a forced-choice questionnaire.

In Table 2, we outline the movements which have the highest agreement between evaluators for each of the four emotions. It shows that by changing the dynamics of the same gesture, the robot can produce recognizable emotions at more than 60% inter-rater agreement. These values are not an exhaustive list of possibilities, but it gives a useful hint for designing motions with these emotions. We also found that the average recognition rates over all samples are significantly greater than chance (25%), suggesting that the DESIRE framework indeed converts the source vocal emotion to the same emotion in gesture.

4 Conclusions and future work

In this study, we verified a hypothesis that emotion from voice could be effectively transferred to motion through only four features (speed, intensity, regularity and extent), giving evidence to our framework for cross-modal emotion analysis and expression. Other future work includes exploring other emotions, mappings to other robots, making the system run online, and integrating other emotional cues such as pose.

This research was partially supported by Kakenhi (S) and GCOE program.

References

[1] M. Farquhar, "Elderly people's definitions of quality of life." *Social Science and Medicine*, vol. 41, no. 10, pp. 1439–1446, 1995.

[2] E. T. Rolls, "Précis of The brain and emotion," *Behavioral and Brain Sciences*, pp. 177–233, 2000.

[3] N. Tooru, M. Taketoshi, and S. Tomomasa, "Quantitative Analysis of Impression of Robot Bodily Expression Based on Laban Movement Theory." *Journal of the Robotics Society of Japan*, vol. 19, no. 2, pp. 252–259, 2001.

[4] C. Breazeal, *Designing sociable robots*, 1st ed. The MIT Press, 2004.

[5] M. Saerbeck and C. Bartneck, "Perception of Affect Elicited by Robot Motion," in *HRI*, pp. 53–60, 2010.

[6] A. Lim, T. Ogata and H. G. Okuno, "Towards expressive musical robots: A cross-modal framework for emotional gesture, voice and music" *EURASIP Journal on Audio, Speech, and Music Processing*, accepted Dec. 6, 2011.

[7] P. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, 2003.

[8] A. Camurri and G. Volpe, "Communicating Expressiveness and Affect in Multimodal Interactive Systems," *IEEE Multimedia*, vol. 12, no. 1, pp. 43–53, 2005.

[9] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[10] P. H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman, and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements," *Revue De Laryngologie - Otologie - Rhinologie*, vol. 117, no. 3, pp. 219–224, 1996.

[11] I. J. Ross, H. L. Shaffer, A. Gohen, R. Freudberg, and H. J. Manley. "Average Magnitude Difference Function Pitch Extractor," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 5, pp. 353–362,1974.

[12] M. Mancini and G. Castellano, "Real-time analysis and synthesis of emotional gesture expressivity," in *Proc. of the Doctoral Consortium of ACII*, 2007.

[13] H. G. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[14] K. Amaya, A. Bruderlin, and T. Calvert, "Emotion from Motion," *Graphics Interface*, pp. 222–229, 1996.

[15] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Journal of Personality*, vol. 82, pp. 51–61, 2001.

[4] http://pascal.kgw.tu-berlin.de/emodb/