Improving social telepresence by converting emotional voice to robot gesture

*Angelica Lim, Tetsuya Ogata, and Hiroshi G. Okuno

1. Introduction

In Japan's aging society and abroad, interviews reveal that the elderly consider communication and social contact as important as their health status [1]. A telepresence robot for social visits to elderly family members would thus be a great way to improve their quality of life. To maximize social communication, the telepresence system should convey emotion as clearly as possible [2]. On the contrary, very few telepresence robots convey emotion through body language, useful especially when the face is not visible.

Implementing emotion-rich telepresence faces at least two challenges: 1) non-discrete emotional states, and 2) limitations of pose-based approaches. Firstly, typical approaches "classify" affect, despite arguments such as Fellous': "Implementing emotions as 'states' fails to capture the way emotions emerge, wax and wane, and subside." [3]. Secondly, conventional robot emotion studies focus on body pose. For example, raised arms indicate surprise [4], but these pose-based approaches are limited: 1) the poses are hand-designed and can become repetitive, and 2) the robot cannot do any other gestures at the same time. For example, "angrily pointing" would not be possible. Approaches such as the use of the Laban Movement analysis [5], conveying emotions through dynamic features like weight, space, and time, are promising ways to counteract these difficulties.

To be pose-independent, we can consider their dynamics of gesture. Recent studies in neuroscience show that movement alone may induce emotion. In [6], it is shown that the observation of angry hand actions recruit the same areas of the brain as when viewing an angry face. In psychology, point-light displays of dancers portraying fear, anger, grief, joy, surprise and disgust are recognized significantly above chance (63%) in [7]. Another study recorded point-light displays of actors performing "drinking and knocking movements" in 10 different emotions [8]. Their results show that the Circumplex affect model's [9] pleasantness and activation dimensions can be recovered in the movements. In summary, not only static pose, but dynamics are important for emotion recognition.

An emotion transfer system should model both emotional input and output on a continuous space. Affect models like the Circumplex affect model have been used in previous work to represent emotion in 2-dimensions [10]. In these approaches, the system designer must map low-level features to the two dimensions of valence (i.e., pleasantness) and arousal (i.e., activation) [11]. It is common to use these models for emotion *generation*, but the problem is that it is not always clear how to map the dimensions for emotion *recognition*. Although we design our system to be flexible to any number of emotional expressions, we limit the present study to verifying that our system can convey four of the basic emotions: happiness, sadness, anger, and fear.

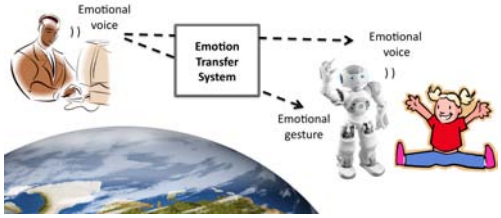

Fig.1: Our emotion transfer system the context of a multimodal telepresence application.

In this paper, we propose an emotional telepresence framework (Fig. 1) to transfer emotional voice to robot gesture, which 1) dispenses with emotion classification, and 2) is pose-independent.

2. An Emotion Transfer Framework

We propose a framework (Fig. 2) 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 [5] [13], and here we study two other parameters called regularity and intensity. Our hypothesis is that SIRE is sufficient for transferring four emotions from voice to gesture. In short, the DESIRE framework is:

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 2: (1) a *hardware-independent layer* mapping DESIRE to perceptual features based on discipline-specific studies in Table 1, and (2) a *hardware-specific layer* mapping the perceptual features to a hardware-specific implementation.

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* [12], pauses [15]	*velocity* [16]
Intensity	gradual vs. abrupt	*voice onset rapidity* [15]	*acceleration* [16], power [18]
Regularity	smooth vs. rough	*jitter* [15], voice quality [12] [15]	*directness* [16], *phase shift* [14] [8]
Extent	small vs. large	*pitch range* [12], loudness [15]	*spatial expansiveness* [18], contraction index [16]

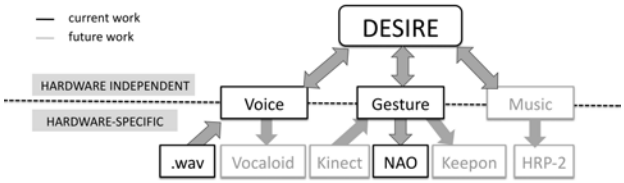

Fig.2: Overview of DESIRE cross-modal emotion transfer framework.

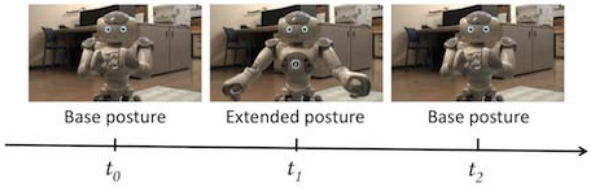

Fig.3: Timeline of an arm gesture.

2·1 Hardware-independent layer

The DESIRE framework was inspired by commonalities found between emotion in movement, voice and music [15] [17]. In these fields, the parameters of speed, intensity, regularity and range are not new, but have been described in varying ways. For example, speed is called *rate* in speech literature [12] or velocity in gesture [16]. We have summarized our literature review in Table 1. This table is not meant to be comprehensive, but rather to give some practical guidelines for how to map SIRE to various modalities.

2·2 Hardware-specific implementation

We provide here the mappings shown in Fig. 2 for 1) extracting SIRE from emotional speech audio samples, and 2) generating motions from SIRE on the NAO Humanoid robot. It has been shown that even robot vacuum cleaners can convey emotions from movement [13]; mappings to other robots will be explored in future work.

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, this results from audio files recorded at 16kHz.

Speed is mapped here to speech rate, or more specifically, syllables per second. One language-agnostic option is to detect speech rate through acoustic features only (without speech recognition), although the state-of-the-art in this problem still has about a 26% error rate [19]. For this reason, we manually provide the number of syllables b for the purposes of this study. 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 [20]. 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[1] implementation of the average magnitude difference function (AMDF) [21] to extract the utterance's F0 trajectory, taking extent as the difference between the lowest F0 and the highest F0.

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], though ideally the framework should be easily applied to other robot hardware. 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 bringing t_0 and t_1 temporally

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

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. In other words, we give more chance to creating a highly irregular movement for low values of 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

Experiment 1: Motion only In this experiment, our goal is to find out whether the robot could produce recognizable emotions through motion and if so, find out what SIRE values can produce each of four emotions. Additionally, by using SIRE values extracted from voice data, we test how reliably our system converts a vocally expressed emotion to the same emotion on a gesturing robot. We recruited 20 normal-sighted evaluators from Kyoto University Graduate School of Informatics. The participants were males of Japanese nationality, ranging in age from 21-61 (mean=27.1, stdev=8.9). As input, we used audio samples taken from the Berlin Database of Emotional Speech[2], which is a database of emotional speech recorded by professional German actors. Each sample was a normalized wave file at 16kHz, 1.5 to 3.9 seconds 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 the 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 used (an extension of both arms in front of the robot, as in Fig. 3), repeated four times in series for each sequence. The sequences were shown in a random order to participants in a classroom. After each sequence, the participants were given 5 seconds to choose 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 the robot, by changing the dynamics of the same gesture, can produce recognizable emotions at more than 60% inter-rater agreement. Table 2 also gives the SIRE parameters which achieve them. These values are not an exhaustive list of possibilities, but it gives a useful hint for designing motions with these emotions.

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

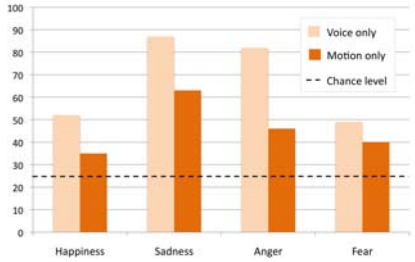

Fig.4: Motion recognition results (Exp. 1) compared to voice only (Exp. 2).

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

Figure 4 shows the aggregated recognition result of each emotion converted from voice to gesture. We find that the recognition rates for all emotions are significantly greater than chance (25%), suggesting that the DESIRE framework indeed converts the source vocal emotion to the same emotion in gesture. We compare the motion recognition results with the voice recognition results from Experiment 2 next, which, as the source input, act as an upper bound.

Experiment 2: Motion and voice In this experiment, we aim to assess the usefulness of the DESIRE system for emotional expression via telepresence. We compare the emotion recognition of 1) a humanoid playing a voice only with 2) a humanoid playing a voice *and* performing a motion created using the DESIRE system. Additionally we test whether the system is effective for more than one gesture. We recruited 21 male evaluators from Kyoto University Graduate School of Informatics ranging in age from 21-27 (mean=24.5, stdev=4.1).

This experiment was performed with a NAO robot placed on a table in front of the participant. The robot was programmed to generate a head movement and a randomly chosen arm gesture (either both arms extending forward, or raising one hand while lowering the other). The gesture dynamics were generated using the SIRE values extracted offline from the 16 utterances described in the Experiment 1.

We presented the participants with two conditions.

- **Condition 1: Voice only**. The robot stayed still in a neutral position while the vocal utterance was played through the 2 speakers in the robot's head.
- **Condition 2: Voice + Motion**. The robot moved according to the SIRE parameters found from the vocal utterance playing simultaneously through its speakers.

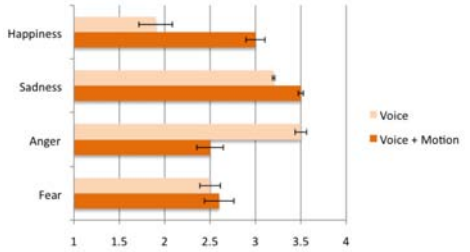

Fig.5: Experiment 2: Comparison of ease of understanding, from difficult (1) to easy (4), for correctly recognized samples.

Each of the 16 utterances were shown in the two conditions in a random order. Evaluators were given 5 seconds after each sequence to choose the one emotion (happiness, sadness, anger, and fear) they thought the robot was conveying the most. Additionally, they rated the difficulty in understanding the robot's conveyed emotion, using a 4-point Lickert scale ranging from "easy to understand" to "hard to understand".

Our results show that for the emotion most difficult to recognize through voice only –fear– the addition of motion increased recognition from 49% to 55%. Next, we compare the evaluator's ratings for "ease of understanding": for a given rater, when a sample was recognized correctly for both voice and voice+motion, we compare the rater's ease of understanding for the two sequences.

In Fig. 5, we notice that anger is better understood when the robot is still than when the robot is moving. This can be due to the choice of "neutral" stance during the voice-only condition; the robot is staring straight forward with hands closed. A maintained stare has been found to be a sign of hostility or anger for both people and animals [22]. On the contrary, the movements generated in our experiment included head movements that turned left and right when regularity was low (R was less than 0.2 in all anger samples). Experiment 1 gives further evidence to this "staring" effect, because recognition of anger was relatively high with arm movements and a still head. This suggests that a humanoid that maintains a forward-facing stare may be viewed as angry, which could have general implications in HRI as to how robots are perceived.

We also see that in Fig. 5 that happiness in particular is more clearly portrayed through voice+motion than through voice only. This may be explained by the fact that the neutral position pose of a stationary robot is quite different from the energetic portrayals of happiness that typically accompany happy voices. This suggests that a gesture may be very useful for portraying joy through a telepresence robot, an important result for our social visit application.

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). Our analysis provided two other surprising results. The first is that happiness is more easily conveyed with motion than with voice alone. Secondly, robots with no head motion can be easily interpreted as angry. This underlines the importance of integrating motion into our telepresence systems.

Our result also shows that the interaction between static poses (such as head pose) and movements should be studied further. Other future work includes making the system run online, integrating other emotional cues such as pose, and choosing how to adapt the system automatically to various users.

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] J. Fellous, "From Human Emotions to Robot Emotions," *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, American Association for Artificial Intelligence*, pp. 39–46, 2004.
[4] M. Zecca, N. Endo, S. Momoki, K. Itoh, A. Takanishi, "Design of the humanoid robot KOBIAN - preliminary analysis of facial and whole body emotion expression capabilities," Humanoids, pp.487–492, 2008.
[5] 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.
[6] C. N. Unit, M. Neurological, and B. Centre, "Brain Networks Involved in Viewing Angry Hands or Faces," *Cerebral Cortex*, vol. 16, no. 8, pp. 1087–1096, 2006.
[7] W. H. Dittrich, T. Troscianko, S. E. Lea, and D. Morgan, "Perception of emotion from dynamic point-light displays represented in dance," *Perception*, vol. 25, no. 6, pp. 727–738, 1996.
[8] 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.
[9] J. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
[10] A. Beck, A. Hiolle, A. Mazel, and R. Losserand, "Interpretation of Emotional Body Language Displayed by Robots," *AFFINE*, pp. 37–42, 2010.
[11] C. Breazeal, *Designing sociable robots*, 1st ed. The MIT Press, May 2004.
[12] 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.
[13] M. Saerbeck and C. Bartneck, "Perception of Affect Elicited by Robot Motion," in *HRI*, pp. 53–60, 2010.
[14] K. Amaya, A. Bruderlin, and T. Calvert, "Emotion from Motion," *Graphics Interface*, pp. 222–229, 1996.
[15] 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.
[16] M. Mancini and G. Castellano, "Real-time analysis and synthesis of emotional gesture expressivity," in *Proc. of the Doctoral Consortium of ACII*, 2007.
[17] A. Camurri and G. Volpe, "Communicating Expressiveness and Affect in Multimodal Interactive Systems," *IEEE Multimedia*, vol. 12, no. 1, pp. 43–53, 2005.
[18] H. G. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, vol. 28, no. 6, pp. 879–896, 1998.
[19] D. Wang, S. S. Narayanan, and S. Member, "Robust Speech Rate Estimation for Spontaneous Speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.
[20] 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.
[21] 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.
[22] P. Ellsworth, J. M. Carlsmith, "Eye contact and gaze aversion in an aggressive encounter" *Journal of Personality and Social Psychology*, vol. 28, no. 2, pp. 280–292, 1973.